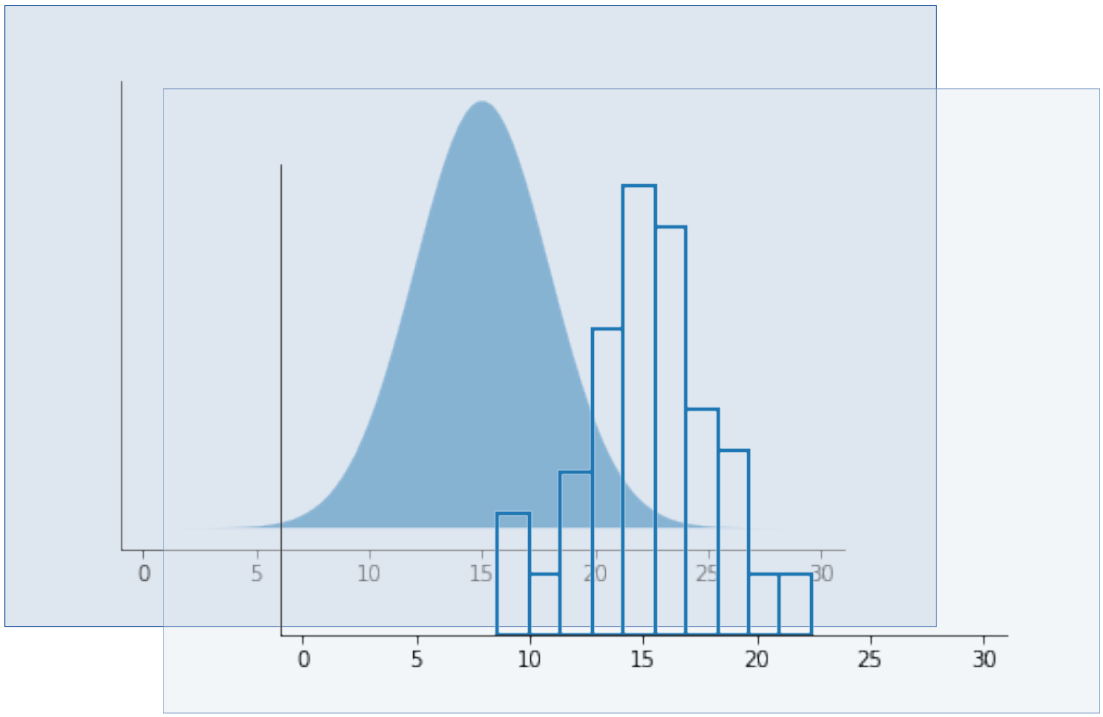
Introduction to statistics with CogStat

Attila Krajcsi



First edition

Publisher Eötvös Loránd University, Faculty of Education and Psychology

elte | ppk

Author Attila Krajcsi

Reviewer Karolina Janacsek

Edition First edition, finalized on April 30, 2020.

ISBN 978-963-489-219-9

DOI 10.21862/978-963-489-219-9

The textbook was supported by the ELTE Eötvös Loránd University textbook and lecture note writer grant.

Table of Contents

1 Preface

This textbook was made for two groups of readers in mind. First, the text is for absolute beginners who are just about to learn statistics and to apply it. Large part of the text explains details that were written for them. Second, a smaller part of the text is for anyone who has already learned statistics, but want to clarify some misunderstandings, want to see some nontrivial or non traditional solutions, want to get more insight how the analyses work.

Accordingly, the text includes two types of sections or parts. The regular text is the larger part of this textbook and it is intended to be read by the beginners. The smaller part of the text is labeled as Advanced, and it was written for readers who has already learned and applied statistics. I recommend for the beginners not to read the advanced parts together with the regular parts, but they should read the whole regular text first, and then, in a second round they could read the advanced parts.

The textbook is written partly supporting the use of the CogStat software. CogStat was designed to make the analysis more simple, and to make the results to be more understandable and easier to interpret. To this aim CogStat uses some unusual methods to present the results and occasionally CogStat applies some unusual methods to analyze the data. The textbook accommodates to CogStat in two ways. First, trivially, the text presents analyses that were run in CogStat. Second, and more importantly, the text wants to highlight the conceptual background of the analyses. Still, technically, the textbook could be used with any other software, but the sections that describe the CogStat specific details have a CogStat label. Additionally, many details of how to use CogStat is included in the documentation of the software, so this textbook is not an extensive description of the software, but it is an introductory text for statistics where basics of CogStat use is also explained.

Labels used in the text (either for the whole chapter or for a single paragraph or for the rest of the paragraph):

- Advanced. Advanced parts regularly rely on concepts that are only explained later in the Beginner parts. For this reason, for beginners it is recommended to read other parts first, and then read the Advanced parts only later.
- Text with CogStat label are specific to that topic.
- Exercise label denotes an exercise.

The description is based on CogStat version 1.9. In newer versions the charts and generally the output might look differently, and some of the analyses might return different details. See the changes since the 1.9 version here.

Acknowledgments I thank Karolina Janacsek for her comments on the text, she was the reviewer of this textbook. I also thank Rebeka Árpási, Letícia Boda, Tibor Csernó, András Csép, Eszter Kiss, Gergely Wintsche for their comments on earlier versions of the text.

2 Measurement levels

Measurement level is one of the properties of the data or variables. It shows how rich information the data include and what mathematical operations can be used with them. Current analyses usually distinguish three types of measurement levels: nominal, ordinal and interval. (There are debates about how data measurement levels should be categorized (e.g., see this Wikipedia article about measurement levels), but in practice (e.g., in statistical software packages) mainly these three levels are used.)

It is essential to learn to identify what type your data is because the available analyses depend on the measurement level of your data, and miscategorization of your data may lead to incorrect calculations and consequently incorrect conclusions.

Nominal data. Nominal data include values that specify membership or classification. In a nominal scale, it only makes sense to tell whether two values are the same or not, but it does not make sense to compare them and to tell which one is larger (they cannot be ordered), or it does not make sense to make arithmetic operations, like addition, multiplication, etc.

For example, gender is a nominal data. Gender can be male or female, and it does not make sense to tell which gender is better or which one has higher values. As another example, in an experiment participants can be enrolled into different groups and the effect of the group manipulation can be studies. In this case, group membership is also a nominal data.

While it can be tempting to think about those values as orders or numbers this would be incorrect and it would be based on some associated properties. For example, there could be gender related properties that may have orders (e.g., men are usually taller, or in most countries men have higher salaries for the same job than women), but the gender itself cannot be ordered. Also note that although in many cases gender is decoded as 1 for men and 2 for women, the number 1 and 2 did not mean to denote the order in this case. In other words, even if numbers could be ordered, here, numbers are simply used as labels, without their numerical meanings. Generally, nominal data can be either numeric or non-numeric (e.g., words), and no matter which form they take, the only meaningful operation you might use is telling whether two values are the same or not, even if numerical labels are used for those values and it might be tempting to make other calculations with those numbers.

Ordinal data. Ordinal data include values that have an order, but they do not include the information how far the neighboring values are from each other. In an ordinal scale one can not only tell if two values are the same or not, but can also compare the values to tell which one is larger or smaller. However, it does not make sense to perform arithmetic operation, like adding or multiplying those numbers.

As an example of ordinal data, in a running event, the order of the runners can tell who was the first and the second in the race, but the order cannot tell what time difference they had. So we cannot tell whether the first runner was 1 sec faster compared to the second runner, or 5 seconds faster, or maybe just 0.01 sec faster. More importantly, for the same reasons we don't know whether the difference between the first and second runners is the same as between the second and third runners. Another example is the completed education stages, e.g., primary school, secondary school, where higher level education means more knowledge, but the amount of

knowledge cannot be quantified, therefore the distance between the education levels cannot be guaranteed to be equal.

Interval data. Interval data include values that can tell the numerical difference of those values. In interval scales, beyond comparison, it makes sense to perform addition and subtraction. Importantly, this measurement level does not have a natural 0 point, and for this reason, it does not make sense to perform multiplication or division. (This latter property might sound a bit unnatural: While these values are numbers, only limited set of operations could be applied. Still, this is a limitation statistical analyses should adhere to otherwise the analyses results would be incorrect.)

For example, the Celsius scale to measure temperature is an interval scale because neighboring values on that scale have a unit difference, but the 0 point is arbitrary in a sense, that a specific state of an arbitrary material serves as a 0 point (i.e., in Celsius scale, zero point is defined as the freezing point of water at 1 standard atmosphere pressure, but other materials or other states of matter could form a usable temperature scale, too). For example, on a Celsius scale, it does not make sense to say that 20 °C today is twice as hot as 10 °C yesterday because the zero point is arbitrary. Although it is tempting to believe that because 20 is twice as large as 10, in fact, this numerical property cannot be transferred to the Celsius scale. To illustrate that point, imagine the same temperatures in Fahrenheit, which are 68 °F and 50 °F, where the ratio is not 200% (as in °C) but 136%. (And also note that because the Fahrenheit scale is also an interval scale, it doesn't make sense to divide those °F values either.)

Advanced While most textbooks explain that one cannot multiply or divide the values of the interval scales, there are formulas used for interval scales that use multiplication and division, such as the standard deviation of a variable. Isn't that incorrect? In fact, it is only the values of those scales that cannot be used directly in the multiplications or divisions, but other values can be used. For example, it makes sense to talk about the ratio of differences in interval scales, e.g., it makes sense to tell that the temperature change (the difference between the lowest and highest temperature) yesterday was twice as large as today. To further illustrate this latter point, if this statement is true in °C, the same is true in °F, too. As another example, it is in fact the difference of the scale values that are used in square and square root calculations in the standard deviation formula.

Ratio scale. Previously we stated that there are three measurement levels that are used in practice, but still here is a fourth one, the ratio scale, that is also frequently mentioned. It will be discussed soon why this frequently introduced measurement scale is hardly used in practice.

The main additional feature of the ratio scales compared to the interval scales is that while interval scales do not include a natural 0 point, ratio scales do. In a ratio scale, all arithmetical operations can be used, so it makes sense to multiply and divide values.

For example, length of a distance is a ratio scale because 0 makes sense as the minimal possible distance. Another example is the Kelvin temperature scale because in the Kelvin scale (unlike in the Celsius and Fahrenheit scales) the 0 point is not an arbitrary point, but it is the smallest possible temperature.

While most textbooks introduce both the interval and the ratio scales, practically ratio scales are handled as interval scales, and in practice no statistical analyses are used that would rely on the extra information a ratio scale includes. Advanced See more details about this possibility and an

explanation why ratio scales can be transformed to interval variables below and in the Changing the measurement level section.

Summary of the measurement levels

Measurement level	Key features	Available operations
Nominal	Different values	=, ≠
Ordinal	Ordered values	=, ≠, <, >
Interval	Equal intervals between values	=, ≠, <, >, +, -
Ratio	Equal intervals between values and a natural zero point	=, ≠, <, >, +, -, ×, /

Table Summary of some properties of the measurement levels

One can consider these four measurement levels as a series of levels, nominal being the lowest and ratio being the highest level, with more and more information included in higher measurement levels, and higher measurement levels allowing more mathematical operations.

Advanced Find additional details about measurement levels in the Changing the measurement level and the Issues with specifying the measurement level sections. See some of the debates about these measurement levels and some alternative solutions in this Wikipedia article. Also see the insightful summary about measurement theory by Warren S. Sarle.

CogStat In CogStat when the data are displayed graphically, the chart axes reflect the measurement levels. See more details in here.

2.1 Changing the measurement level

Advanced

Measurement levels can be considered as a series or hierarchy of levels, and higher-level data include more information. Based on this order, it is possible to transform the highest available measurement level of the data to a lower level. (See below the possibility of whether low-level data can be transformed to a higher measurement level.) For example, one can consider running time in a race as a ratio scale, then calculating the ranks of the participants as an ordinal scale, or telling if any two participants ended the race at the same time (or in other words, what racers were ties) to get the nominal data. Obviously, the lower measurement level is used, the less information is kept. In other words, when a data with a specific measurement level is transformed to a lower measurement level, we lose some of the information. Transformation to a lower measurement level also means that we drop some part of the information so that only mathematical operations appropriate for lower measurement level will be available.

Note that in practical analyses, technically one shouldn't transform the values of the data to handle the data as a lower measurement level, for example, interval data are not required to be transformed to rank data. Instead, one should just consider the data as a lower measurement data, and apply statistical analyses that are appropriate for those lower measurement levels. For example, if you want to use an interval variable as an ordinal variable, instead of calculating the mean, you can calculate the median, or instead of using a one-sample t-test, you can use the Wilcoxon signed-rank test. In other words, one can use analyses that utilize e.g., only the ordinal or

nominal information in those data. CogStat Similarly, in CogStat one should only change the measurement level in the source data (e.g., change from interval to ordinal), without changing anything in the data values themselves, and CogStat will rely on the appropriate information in line with that data type (e.g., for ordinal variable only the order of the data will be considered, but not the specific values). See section Handling data in CogStat how to set the measurement level for an analysis in CogStat.

This possible transformation also means that not only statistical methods for a specific measurement level can be used for a data set, but also any statistical methods that are appropriate for a lower measurement level. For example, for an ordinal variable not only the methods for ordinal variables can be used, but also any other methods for nominal variables. In other word, a higher measurement level data includes all the information that would be available if the same properties were measured with a lower measurement level scale.

But why would one want to transform the data to a lower level, when information would be lost? In fact, there are various reasons to do it. (1) For example, in practice, there are only a few analyses that would rely on the extra features a ratios scale includes compared to interval scale features (e.g., geometric mean or coefficient of variation requires ratio scale), while most of the important and common statistical questions can still be answered appropriately with the interval scale methods. Therefore, most of the time ratio scales are simply handled as an interval scale, even if in this case we lose some information. (2) As another example, we might want to make our analysis comparable with some former analyses, and we'd want to handle our higher measurement level data as the lower measurement level analyses of former data. (3) Yet another example is that in many cases, the lower level statistics is more robust, e.g., unlike Pearson correlation, Spearman correlation is not sensitive to outliers, or similarly, median is less sensitive to outliers than mean is. CogStat This latter example also shows why CogStat occasionally calculates some results that belong to lower measurement levels when that result might be relevant. For example, a median is calculated not only for ordinal but also for interval variables, or Spearman correlation is calculated also for both ordinal and interval variables.

Still, in most cases, we want to keep the highest measurement level for our data that the actual measurement method allows because that's how we can keep the maximal information in our data, and usually the highest sensitivity of our analysis.

Usually, it is not possible to convert a measurement level to a higher one than its original highest measurement level because we would add some false information that was not included in the measured data. For example, if one handles the order of the race as an interval scale, one should suppose that the participants arrived with equal time differences to the end, which is most probably was not the case, so this false presupposition would add false information to our data. However, there could be some rare and special exceptions, when data can be transformed to a higher level. See some examples in the summary of Warren S. Sarle.

2.2 Issues with specifying the measurement level

Advanced

While seemingly it is simple to tell what measurement level a variable has, there could be various difficulties, and specifying the measurement level could be fuzzy.

2.2.1 Some example problems

One frequently cited issue is the measurement level of a Likert scale, such as the possible never-rarely-usually-always responses in a survey. Strictly speaking, the distances between the levels are not equal, or at least one cannot guarantee that participants filling out the survey will handle the responses in that way. Therefore, a Likert scale should be an ordinal scale. On the other hand, many times it is found that those levels have approximately equal distances from the neighboring levels, so it can be handled as an interval scale. As a compromise, sometimes Likert scale is characterized as something in between ordinal and interval scales because the distance between the neighboring values is approximately equal, i.e., the distances are neither perfectly equal nor entirely arbitrary.

Another example is the measurement level of reaction time in behavioral studies. Time in itself should be a ratio scale because the scale includes regular intervals, and it has a natural zero point. The problem is that we are not interested in the time in itself, but the performance of the participants reflected in the reaction time. When the reaction time is considered as the performance of the participant, there are several issues with the ratio scale interpretation. First, the natural zero point of a performance index would be the point where the participant cannot solve the task. However, in the reaction time scale, this natural zero point is not the zero value, but some very large value, and the values close to the reaction time scale zero point are in fact the maximal performance values. In other words, the reaction time as a performance index is reversed compared to the reaction time as time measurement index. Second, the maximum performance value is not 0 millisecond because a zero ms reaction time is impossible, but a few milliseconds minimum processing time is required, although it is impossible to say what the minimum value really is. (For example, although the time limit is somewhat debated, in athletics if someone starts earlier than 100 ms after firing the starter's gun, it is considered as a false start.) So far, the reaction time as a performance index is clearly not a ratio scale. Third, as a performance index, it is not guaranteed that any two neighboring values have the same distance. So it is possible that the reaction time is not even an interval scale. (See the diffusion model how in some cases direct use of the reaction time data can be got around.)

A third problematic example is the error rate in behavioral studies. Many times, error rate (e.g., 12% of the trials/tasks are solved incorrectly) is handled as a ratio scale, and as a measurement of correctly (or erroneously) solved tasks this could be true (because it has a natural zero point, and the differences of neighboring values are equal). However, as a measurement of performance level, there is no guarantee that the difficulties of the tasks are uniformly distributed, or that some other properties of the task will warrant that the neighboring values of the error rate have the same distances. (See also the Rasch model that discusses this problem, and see the diffusion model that handles some of the tasks with error rates together with the reaction time data.)

2.2.2 Solutions

How could these issues be solved? There are various considerations that can help to decide in those cases.

1. One of the most important things is that the measurement level should be considered as the measurement level of the conceptual thing one wants to investigate. For example, the reaction time can be considered both as time measurement and as performance measurement. Clearly, in

behavioral studies, the reaction time is an indirect performance measurement, and it cannot be a ratio scale.

2. Another viewpoint is that in some cases it does not matter which measurement level we choose because either way the same analysis will be run. For example, usually, the question is whether a scale is ordinal or interval. However, many hypothesis tests for interval scales also require that the data should be normally distributed. When an interval variable is not normally distributed, many times a hypothesis test is chosen that is otherwise used for ordinal variables. In those cases, no matter if we decide to have an ordinal or interval variables, in the end, the same tests will be run.

3. Additionally, even if some measurements are ordinal, many times we are not using the items directly, but some aggregated index of several items, such as the mean of some Likert scale items. Summary of random variables tend to be normal, and in these cases, it is not unreasonable to handle these indexes as interval variables.

4. There could be additional considerations, that might help to tell more reliably what measurement scales the data have. See various publications about measurement levels for more details, such as the example of Rasch model or the diffusion models.

5. Finally, when in doubt, it is safer to choose the lower measurement level, so a questionable part of the information will not be utilized.

3 Handling data in CogStat

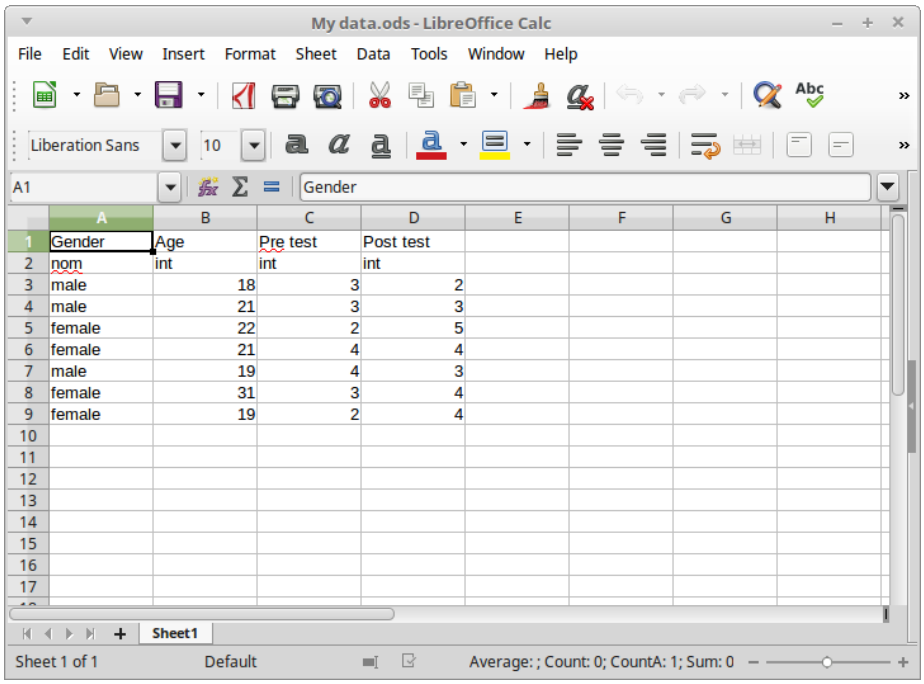
CogStat

Like in most statistical software packages, in CogStat, data are stored in a table, where rows are the cases, such as participants (i.e., the data in that row belongs to a specific participant of the study, and all of the data of that participant is included in that single row; see the example table in the Figure below) or trials (i.e., the details of a single short task is included in a row, such as the identifier of the participant, the stimuli that has been shown, the response the participant gave, the reaction time, etc.), and columns are the variables (such as the age of a participant, or the stimulus that was shown in a trial, or a response to a specific item in a test, etc.). Also, like in most statistical software, the variables have a name with which we can refer to those variables in the analysis and in the results. Finally, unlike in most statistical software packages, in CogStat you should set the measurement level of your variables. This is essential because CogStat will choose automatically the most appropriate analyses partly based on this information. If you set the measurement level incorrectly, CogStat might give you inaccurate or improper results.

Unlike most statistical software, CogStat does not have its own data handling interface, but you should rely on other software: You'd need either a spreadsheet software or you need SPSS. Still, it is very easy to import the data, and this extra data importing step does not require much effort.

If you store your data in SPSS (or any other free compatible alternative, such as the PSPP), remember to set the measurement level of your variables and to save the file before importing them.

Also, you can store your data in a spreadsheet software, such as LibreOffice Calc or Microsoft Office Excel. See more details here and here.



	A	B	C	D
1	Gender	Age	Pre test	Post test
2	nom	int	int	int
3	male	18	3	2
4	male	21	3	3
5	female	22	2	5
6	female	21	4	4
7	male	19	4	3
8	female	31	3	4
9	female	19	2	4

Figure Data in LibreOffice Calc

In your spreadsheet, your data should include the following pieces of information (see more details here):

1. In the first row, store the names of your variables.

2. In the second row, store the measurement level of your variables. This could be either ‘nom’, ‘ord’ or ‘int’ for nominal, ordinal or interval variables, respectively. (See also the section Measurement levels.)

3. From the third row, store the values of your variables. Each row is a single case (e.g., a case includes the data of a participant, or the data of a trial in an experiment, etc.).

You can either save your data as .csv file (make sure you use tab and not comma as your field separator) and open it in CogStat, or you can copy and paste your data from your spreadsheet to CogStat (make sure that in your spreadsheet software the decimal separator is set to dot to be in line with the CogStat settings.). We recommend to use the copy and paste method. (Again, see more details here and here.)

4 Samples and populations

Usually, when measuring something, we are not simply interested in the specific cases we've just measured, but we want to say something more general about the larger group the measured cases were taken from. For example, in research, if we measure the symptoms of a group who took some medication and a group who received placebo, we are not only interested in the symptoms in those specific persons, but more generally, we want to know whether the medication could be effective for anyone. In statistics, we clearly differentiate these two types of groups.

The **sample** is the group of cases we have measured, and the **population** is the larger group we want to generalize our measured results to.

There are exceptional cases, when the whole population could be measured. In those cases, our sample is the population as well. But this is quite rare, and in almost all cases what we measure is only some smaller part of the whole population.

So while usually we're interested in the population, we cannot measure it (because it would require too many resources, part of them is unavailable, etc.). And while we measure only the sample, we want to learn about the whole population.

One main aim of the statistical analysis is to say something about the population when only a sample is available. Statistics provides many methods with which one can make statements about the properties of the whole population based on the sample data.

Related to the differentiation between the samples and the populations, mathematical statistics can be divided into descriptive statistics (mostly dealing with the sample), and inferential statistics (reasoning about the population based on the sample data). **Descriptive statistics** summarizes some properties of the sample. (If we measured the whole population and our sample is also our population, then descriptive statistics summarizes the population.) **Inferential statistics** tries to find the population properties when the whole population could not be measured and only a part of it (i.e., the sample) is known.

The general term for the calculated statistical properties of the data, such as the mean, are also different for the actually measured data (sample) and for the inferred properties (population). Calculated statistical property of the sample (or the population when the whole population is measured) is called the **descriptive statistic** (or simply, statistic, and sometimes descriptive), while the inferred property of the population is called the **parameter**.

Part of mathematical statistics:	**Descriptive statistics**	**Inferential statistics**
Terms for the properties:	Descriptive statistic, statistic or descriptive	Parameter
Role in practical work:	Summarizes some properties of the sample, or when the whole population is measured, summarizes the population	Infers the properties of the population based on the sample

Table Related terms and role of descriptive and inferential statistics

CogStat In most analyses, CogStat displays results about the sample and the results about the population separately.

Advanced In many cases, it might seem less straightforward what the population is, or why a sample is appropriate to represent a population. For example, in behavioral studies, while one wants to investigate the human race, usually only psychology university students are measured. In many textbooks, the same psychology student sample is introduced as a sample for the university students population. So do university student participants form samples for university student population or samples for the population of the human race? First, it depends on the intention of the researchers. They can specify what population they are interested in, and then the measured sample can be considered as a sample of that population. Second, when one can suppose that a smaller subpopulation can be representative of the whole population the researcher intends to investigate, the sample could come from that subpopulation. For example, although university students are generally smarter, better educated, and younger than the whole population of humans, it is supposed that many details of the mental processing is universal, e.g., large part of visual processing works similarly in young and older people, so choosing a sample from the subpopulation of university students can be representative for the larger population of humans. It is possible that later it can be found that the subpopulation is not representative to the larger population. In those cases, the conclusions of the former works have to be reconsidered, and additional measurements are required to find out how things work in other segments of the larger population.

4.1 The relation of the samples and the populations

To understand how statistical methods can infer from the sample about the population, we should understand how the sample and the population are related. First, we have to understand what properties a sample may have when the properties of the population are known. To build up some intuition about it, let us run some simulation.

In the following simulation we use a spreadsheet software. Spreadsheet software packages can be used not only for simple calculations and creating charts, but they know a lot of statistical calculations as well, and because they can generate various types of random numbers, they can be used for statistical simulations, too. For the simulation task see “Random sampling” sheet in the related spreadsheet file. We suggest to use LibreOffice for opening and using the file.

For the simulation let’s suppose that we are interested in the height of a population. Let’s also suppose that we know this population: Let’s suppose that it has a normal distribution with known parameters.

Probability distribution and normal distribution. Before moving on, let us clarify what a distribution is, what a probability distribution is, what a normal distribution is, and finally, what the parameters of the normal distributions are? If we measure a property of a sample, the different possible values have different frequencies, which frequency of all measured values form the distribution. In the figure below a test score was measured for 100 participants, and the distribution of the data are shown in a bar chart, where the x-axis shows the score ranges, and the y-axis shows the number of participants having the score in those specific ranges. In other cases we are considering not the specific sample, but the population, which population may potentially have

infinite cases. Even in the case of a population with infinite cases one might think of the distribution of the values, although it is not the number of specific values, but it is a function describing the probabilities of having the specific values. So this function is the **probability distribution** which describes the chance of the a specific value will be be found or measured. Usually, probability distribution is depicted as a line chart (see some examples below; don't worry about what Landau distribution is, it just an example of distribution for us now; the her example, the normal distribution will be explained below), where on the x-axis there are the possible values one can measure, and on the y-axis the probability of measuring that value. See some distribution examples in the figure below.

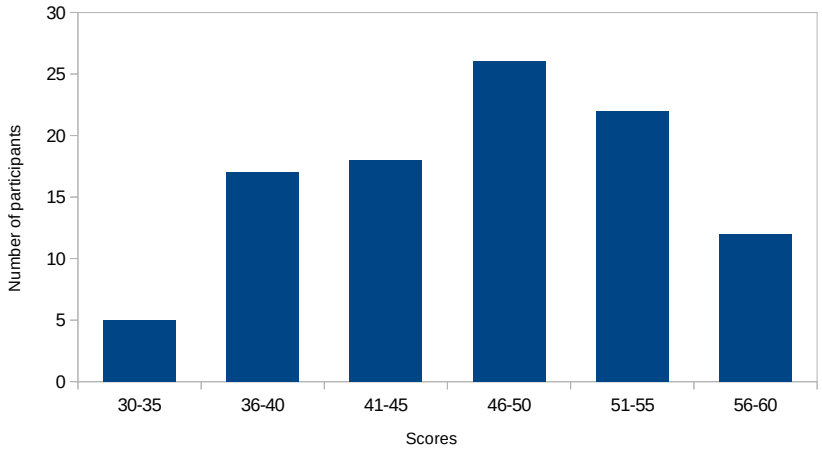


Figure Distribution of a variable measured in a sample

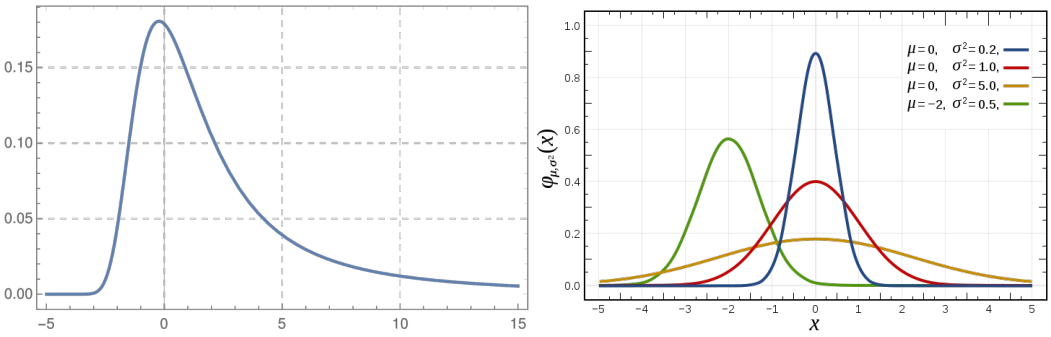


Figure Landau (left, source) and normal (right, source) distributions. In the normal distribution, μ denotes the mean, and σ denotes the standard deviation, and its square is the variance

There are some special types of distributions that can be found frequently. For example, in psychology, biology, and in a series of other areas, data are often have normal distribution. (Advanced See the mathematical reasons why this is seen frequently.) **Normal distribution** (see some examples of them in the figure above) is a type of distribution where there is a typical value, and the farther we get from this typical value, the less likely that we will measure it. The relation of the value and the probability to find it is described by a very specific function, so not all distributions that have a typical value and a decreasing chance with values far from the typical values are normal distributions, but only the ones that take this specifically shaped function. (See the exact function of the normal distribution here to acknowledge that math sometimes can be mysteriously

complicated. For understanding the rest of this description, you don't have to understand how this formula creates this shape.) For a normal distribution, there are two parameters (specific values) that can influence how this distribution exactly looks like. One of them is the typical value, termed the location or the mean, that will determine where the highest point (the most probable value) of this distribution is, or in other words, where the distribution is located on the x-axis (i.e., it shifts the distribution in the left-right direction). The other parameter, termed the variance, specifies how much the distribution is spread out: normal distributions with higher variance are wider, while normal distributions with lower variance are narrower. If the mean and the variance is known for a normal distribution, then everything is known about it (e.g., the curve can be drawn precisely). See the examples again on the figure above how those values influence the distribution.

Getting back to our height example, we suppose that our population has a normal distribution, and let's suppose that the mean is 170, and the standard deviation is 16. (The standard deviation (SD) is also an index to express the spread of a distribution, just like the variance. In fact, the standard deviation is the square root of the variance (or in other words, the variance is the square of the standard deviation). So in this normal distribution example, if the standard deviation is 16, then the variance is 256.) Our population distribution looks like this, and this is what you can see in the spreadsheet on the left:

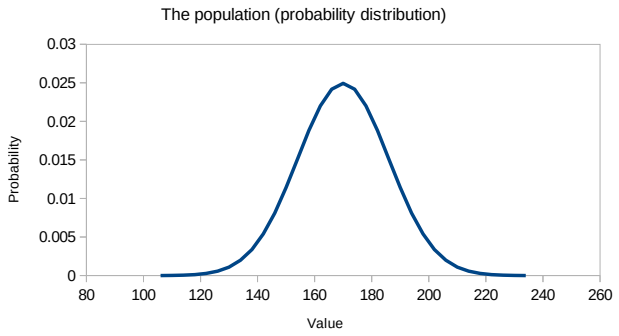


Figure A hypothetical population

The next step in the simulation is sampling. Sampling is the process of choosing cases from that population randomly. (Technically, in the spreadsheet, we generate random numbers from the supposed normal distribution. Random number from a specific distribution means that selecting a specific number has the same chance as it is described in the probability distribution. This also means that for the long term, the ratio of a specific value in the larger sample is the same as the probability of that value in the population distribution.) In the present example, 15 cases were sampled. This can be considered as a measurement, where 15 cases were measured from the population. The data are also displayed in a dot plot (see an example in the Figure below). In a **dot plot**, every single dot is a single case, and the vertical position of the dot shows the value of that case. To get another type of impression about the data, the histogram is also displayed (see an example in the Figure below). In a **histogram**, the whole data range is divided into equal width parts, and the number of cases in every single part is counted and displayed as the height of the bars. Histogram is a type of summary of the data, and it displays in which range there are more cases, and in which range there are less. In other words, the histogram displays the distribution of the sample. The histogram of our sample can be compared to our expected population distribution.

On our spreadsheet, the mean and the standard deviation of the sample is also calculated. This can also be compared to the population mean and population standard deviation.

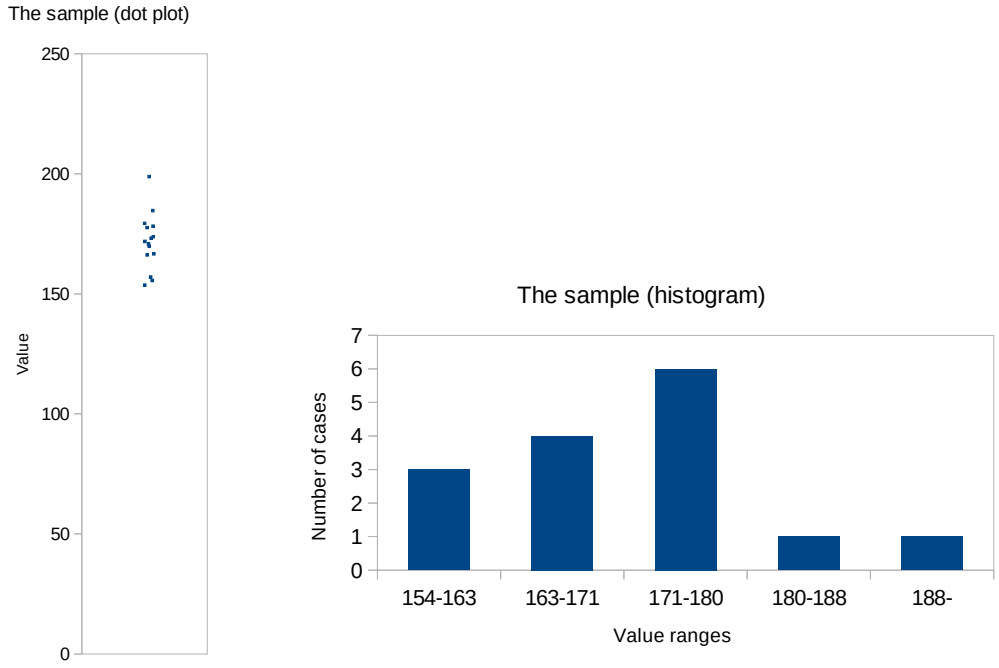


Figure A sample from the hypothetical distribution

Now an essential part of building up some intuition about how the sample and the population are related is to repeat the sampling (creating additional samples from that population) and checking how the samples are changing. On the spreadsheet, go to an empty cell, and press the delete button. Every time you press the delete button, a new series of random numbers are generated. You can imagine it as collecting a new sample from the same population. After every sampling, you can check how similar the sample (the dot plot and the histogram) and the population (the probability distribution) are. First of all, usually, the histogram has different shape compared to the normal distribution of the population, and it seems really random where those specific 15 numbers come from, so the sample and the population are quite different. Also, you might observe that the sample mean and the sample SD (standard deviation) practically never match the population mean and population SD. On the other hand, while the sample is different from the population, it never deviates too radically from the population. For example, the sample mean and the sample SD are moving somewhere around the population values.

As a next step in forming some intuition how sampling works, let's modify the population. Change the mean of the population, and generate again a few more samples. Check the properties of the samples, and compare it with the population properties. You'll see that the probable sample means are shifted in line with the population means. And again you will see some differences between the sample and the population, but again, these differences are somewhat limited.

Importantly, after a while you might have an impression about the probable sample means and sample SDs when the sample comes from that predefined population. Now you might use this knowledge to infer backward: If you see a new sample coming from an unknown population, do you think it might come from that specific population you've been playing with? Imagine you're running a real study where a sample is measured, and the population is not known. The

measurement will not be repeated several times, there is only a single sample. You find that your sample mean is 168 and the SD is 20.4. What could you say about the population? Is it possible that the population mean is 170? It seems reasonable because a population with a mean of 170 quite often produces a 168-mean sample. Is it possible that the population mean is 171? Still, it seems quite reasonable. Now is it possible that our population has a mean of 110? Probably not because that population mean rarely generates a sample mean of 168. These population examples were relatively easy to evaluate because based on our previous experience the population means were relatively close (170 and 171) or relatively far (110) from the example sample. But for some population parameter in between, it would be much harder to guess whether the measured sample could come from that population. For example, could the sample with 168 mean come from a population with a mean of 150? Or what about 155? Still, based on our previous experimenting, we might specify a potential range of population means that are relatively likely values that can generate our measured sample, and practically the farther we go away from that range the less likely it is that our sample is generated from such a population.

So after some experimenting, one can have some experience what samples are typical for a given population, and what samples are not. Based on this knowledge, if a sample is measured, one can estimate what populations are likely to produce that sample, and what populations are unlikely to create a sample like that.

Now that is exactly what inferential statistics does: Based on the rules how samples can be chosen from various populations, it tries to characterize the population with the given sample. In other words: given the sample, inferential statistics tries to find the properties of the underlying population. Inferential statistic has a great advantage over our intuitive guessing: It quantifies the probabilities. Inferential statistic doesn't just tell us if e.g., a sample is likely or unlikely to come from a specific population, but it tells us the exact probabilities.

4.2 Random sampling

To make valid statements about the population, in all statistical analyses, it is essential to use **random sampling**, i.e., all members of the population should have the same chance to be included in the sample. Otherwise, the sample will not be representative of the population, and the inferences about the population will be incorrect. For example, if we measure political views of citizens of a country with an online survey, part of the population who use the internet less frequently will have smaller chance to be included in the sample, and those parts of the population will be underrepresented, and the other parts of the population will be estimated, and the other parts of the population will be overrepresented, resulting in biased conclusions about the population. So representative sample is essential because biased samples will cause biased statistical inferences about the population.

Advanced There are some cases when the representativity of the sample is not strictly required. For example, if one wants to measure whether there is a correlation between two variables or not, some violations to representativeness most probably will not make correlation disappear: for example, mean values have smaller effect on the correlation than more extreme values, so removing cases with mean values is less harmful than removing cases with more extreme values. However, in some other cases representativeness is essential: in the case of correlation if someone wants to know the exact correlational value, non-representative sample will introduce a

bias. It is important to highlight that while some researchers criticize any non-representative sample, in fact, representativeness is not always essential, and one should consider whether non-representative sample may introduce an essential bias in results or the bias is negligible.

Advanced There are cases when the bias is known, and in those cases the biased sampling can be corrected. See some more details here.

4.3 The ultimate limitation of inferential statistic

You can never trust inferential statistic 100%! The simple simulation above can already demonstrate what one of the most important problems is with inferential statistic.

While getting more and more samples in the former simulation we could see that some sample statistics were quite far from the population parameters. Taking the example of the mean, in an extreme situation, one can imagine that all of the cases come from some very unlikely part of the distribution. Mathematically, the normal distribution can include infinitely small and infinitely large values, so theoretically it is possible that our sample includes only very-very small or only very-very large numbers, leading to extremely large or extremely small means. Generally, no matter what population distribution we have, it is possible that in our sample we have only very small or very large values from that distribution, and the sample mean can take any value from the possible values from that population distribution. (In practice, all measurement methods have some bounds, for example a test has a minimum and maximum value, therefore the mean one might measure should be in that interval, but the same idea also applies there, and the mean (or any other descriptive) could be anything within that range.) If we know that a sample could be extreme and we use this knowledge to infer about the population the sample might com from, then we must think that a specific sample can come from any population whose distribution would allow that sample, no matter how unlikely or extreme that sample is. For example, a sample includes the values of 12, 52, 104 and 37. Let's also suppose that we know that this sample is coming from a normal distribution, but we don't know what mean and standard deviation that normal distribution it has. Now if we want to infer about the population mean based on the sample data, then theoretically it is possible that that sample could come from a normal population with any means: Maybe that sample comes from a normal distribution with a mean of 10, or a mean of 150 or a mean of -200. Even if the latter population means are unlikely (because the sample data are rather far away from the population mean, i.e., the sample data would not be typical in those populations), they are still possible (because a normal distribution population mathematically may include any values, even if the extreme values are unlikely). (Note again that in practical measurements there are some limitations what value might measure, e.g., reaction time cannot be a negative number, but this reasoning is also valid in those limited ranges.) In other words,no matter what mean parameter we suppose for the normal distribution population, it is still possible that our sample came from it. This means that if we want to be 100% sure that our list includes the real population that generated our sample and if we want to include all possible situations that could generate the given sample and , then we will get the list of all possible normally distributed populations that might exist. Now this means that our sample could come from any population, therefore we do not know anything about the population, because anything is possible. In other words, if we measure a sample, then our 100% certain conclusion about the population should be that that sample could come from any population. Clearly, this conclusion is not informative at all.

So if we want to be 100% sure that our list of potential populations includes the real one, then our list will include all possible populations, which result is useless. To have a more reasonable solution, we usually don't consider extremely unlikely populations even if technically they cannot be excluded. So in practice we only consider the most probable populations, and by convention psychologists consider the populations that would cover 95% of the most probable cases, and they ignore the rest 5% of them. (Later, more specific description will be given what 95% means in some specific calculations or in some specific statistical methods.) Still, this means that if we don't want to see the useless 100% sure solution, then we'll ignore 5% of the cases that are the most unlikely ones, but that are still possible. This 95% method solves the usability problem of the 100% certain solution, but introduces another crucial problem: Because it is always possible that your sample in fact came from a population that is unlikely (i.e., your sample is extreme) and which population is included in the ignored 5% of populations: you can never trust inferential statistic 100%. This is a key component why inferential statistic is difficult to use and interpret, and unfortunately this is an issue you cannot resolve it any way, but you have to live with it. The root of this issue is that sampling is a random process and while inferential statistics tries to identify the set or range of populations the sample came from, it cannot be 100% sure (or when you force a 100% sure answer, the answer is useless). Again, the bottom line here is that you can never trust inferential statistics 100%, and you have to live with this uncertainty. (See some examples of this issue in parameter estimation or in hypothesis test.)

5 Investigating the sample and the population

When a sample is measured, we want to investigate its properties, and then usually we want to conclude about the population. Here we review the main features of investigating the sample and the population.

5.1 Investigating the sample: Raw data and sample properties

When investigating our data, one essential task is to understand the properties of our sample, i.e., the properties of the data of the actual cases, independent of how these properties can be generalized to the whole population (supposing again that only a small part of the population is measured - in practice almost exclusively this is the case).

Raw data. First, one can directly look at the raw data, without any further calculations, which is usually some kind of chart displaying the individual data. For example, see the individual data of three groups in the figure below. There are at least two reasons why one would want to do this. First, sometimes the response to our research question can already be seen in our raw data because the response is obvious. For example, in some cases, like in the figure below one can see that the values of cases in some groups clearly differ. Or contrariwise, in other cases, one can clearly see that there is no difference at all between the groups. Still, in some other cases, the difference could be in-between, and only by looking at the raw data it is impossible to tell whether the investigated effect exists or not. Still, for the sake of the cases when the result is straightforward, it is worth to look at the raw data. Advanced Second, sometimes the data include some unusual or unexpected pattern the sample statistics (see also below) cannot reveal, and for this reason, the usual sample descriptives or population estimations or hypothesis test results could be misleading. For example, a few outliers (cases with extreme values) might considerably influence our result, while these occasional values are not always representative, and it would be more appropriate to ignore them. As another example, a distribution might be bimodal, when one would expect a unimodal distribution and would interpret the statistical analysis accordingly. More generally, there are cases, when different raw data patterns show the same descriptive statistics. See more details about it here. In those cases, the raw data include useful extra information compared to the sample descriptives.

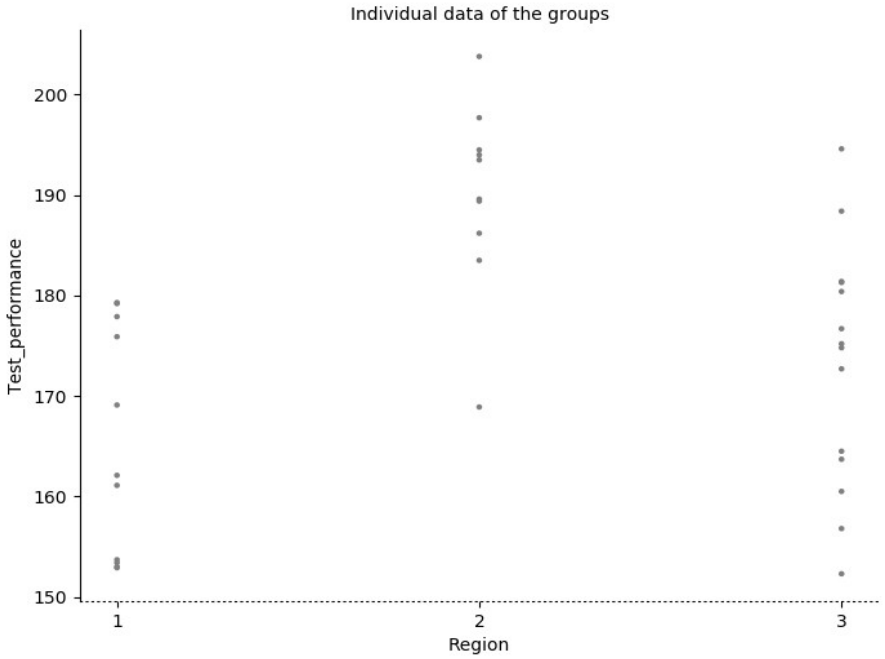


Figure Raw data of three groups

Sample properties. Beyond the raw data, one can calculate indexes that can summarize and characterize the sample data, such as mean, standard deviation, standardized effect sizes, etc. For example, in the figure below, the means clearly set the location of the values for the three groups. These indexes are termed **descriptive statistics** or simply **statistics**. (See more details about the descriptive vs. inferential statistics and the related sample vs. population properties in Samples and populations) There are at least two reasons why this could be useful. First, many times the data of a sample are redundant and can be described in a more compact form. For example, samples often form normal distribution, and normal distribution can be described with two parameters: with its mean and its standard deviation. If one knows those two parameters, then everything is known about that distribution. This also means that the long series of raw data are redundant, and to describe the data more briefly, one can simply use the mean and the standard deviation. Thus, sample properties can describe the data more briefly and more generally. Second, there could be properties of the sample that we cannot sense precisely only by looking at the raw data. For example, the correlation of a data set is frequently hard to see if the correlation is low. (See more details about the correlation in the Linear regression and correlation part.) Or it is hard to tell whether a distribution has either a small standard deviation or a positive kurtosis – see more details about this example in the Skewness and kurtosis section. So, descriptive statistics can find a more precise specific value for a feature.

Sample properties
Descriptives for the groups

	1	2	3
Mean	165.24	190.11	173.09
Standard deviation	10.80	8.95	11.73
Maximum	179.30	203.80	194.60
Upper quartile	176.90	194.38	181.08
Median	162.10	191.55	175.00
Lower quartile	153.55	187.00	163.90
Minimum	152.90	168.90	152.30

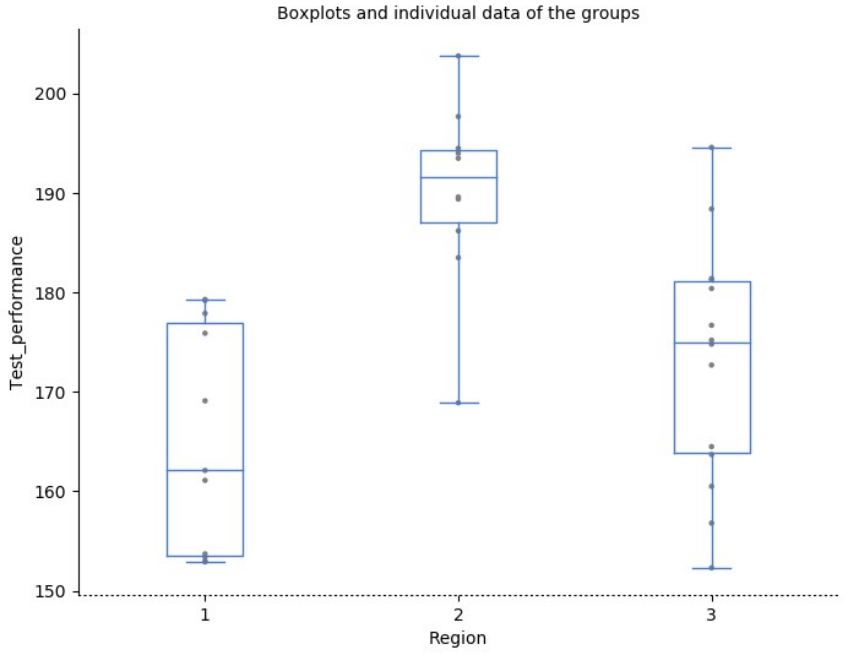


Figure Descriptive data of three groups

Overall, sample descriptives summarize and highlight some critical properties. Still, at the same time some of the details are also hidden, because different kind of raw data may have the same descriptives. Some of those hidden details can be observed by looking at the raw data. For these reasons, it is a good practice to look at both the raw data and the sample properties.

5.2 Investigating the population

The measured sample is only a part of the population, and researchers usually are interested in the properties of the population. There are various methods with which the population properties can be found. We discuss the two most popular methods: parameter estimations and hypothesis tests. (Note that parameters are the various indexes or properties of the population, while the same term for the sample is descriptive statistics.) (Note again that if the whole population is measured, then the descriptives introduced in the previous part should be used, and the whole inferential statistics part is irrelevant, therefore, it can be ignored.)

5.2.1 Investigating the population: Parameter estimation

When the parameters of the population are being estimated, the two most common methods are **point estimation** and **interval estimation**.

Point estimation. With the point estimation, based on the sample, the most probable single value of that parameter is calculated. CogStat In the example of the three groups comparison, the estimated group means are shown in the point estimation table, and in the bar chart (see figure below).

Population properties

Means
Present confidence interval values suppose normality.

Region	Point estimation	95% CI (low)	95% CI (high)
2	190.110000	183.363321	196.856679
3	173.092857	166.065792	180.119923
1	165.236364	157.624362	172.848366

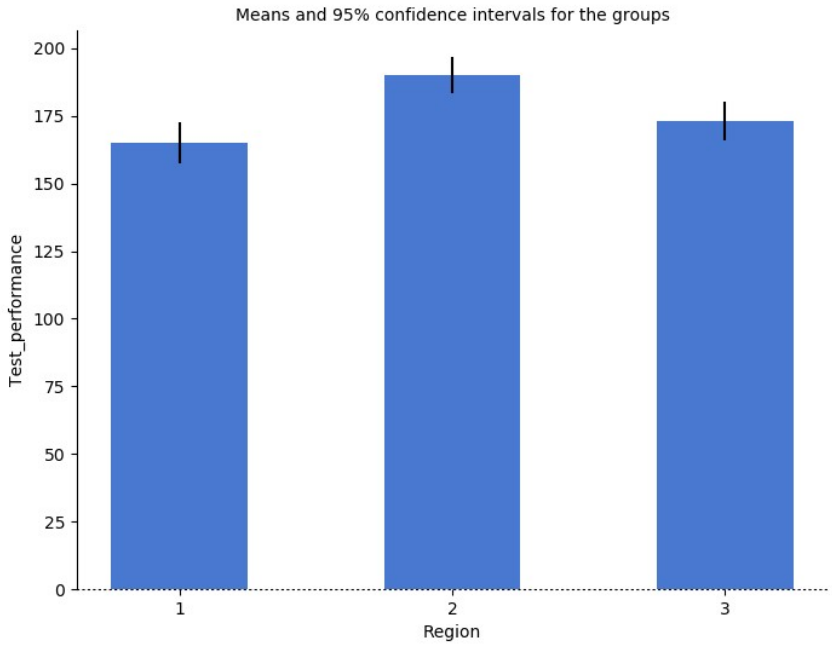


Figure Population parameter estimations of three groups. The top part of the columns represent the point estimation and the error bars (black lines at the top of the columns) represent the confidence interval.

In some cases, the point estimation and the sample statistics are the same values. For example, the point estimation of the mean is the same as the sample mean (see also the sample descriptives and the point estimation in figures of the group comparison example). In other words, the best guess to find out the population mean is to use the sample mean. In some other cases, the point estimation and the sample statistics are not the same. For example, the standard deviation (SD) uses a different formula to calculate the sample SD and the population SD estimation: the population SD is a bit larger than the sample SD, and the smaller the sample is the larger the difference is. (See the mathematical background of it here.)

Interval estimation. With interval estimation, a range of values is calculated in a way that there are great chances that the population parameter is included in that interval. For example, the figure above includes not only the point estimation, but the interval estimation, too. Technically, there are various interval estimation types. Here, our examples will use the confidence interval, the most frequently used interval estimation.

Generally, we could set how sure we want to be that the interval includes the real population parameters. By convention the certainty is set to 95%, and most software provides the 95% interval estimation. Note again that if one wants to be 100% sure, then the interval estimation will be pointlessly wide, so a non 100% value should be chosen. This also means that because the certainty is smaller than 100% (usually 95%), the parameter estimations will never be precise in a sense that we can never be sure that the interval estimation includes the real parameter. Again, this is one of the most important issues in statistical inferences: No matter how sophisticated methods we apply, there will be always a chance that our conclusion is wrong.

Interval estimation is more useful than point estimation. First of all, it is very unlikely that a point estimation really finds the parameter of the population, so it is more appropriate to find an interval which most probably includes the parameter of the population. Second, interval estimation also gives a range not only a value, and the size of the range (i.e., the width of the interval) tells us the precision of our estimation. The narrower the range of the interval is, the more precise our measurement is in a sense that the parameter can be found or estimated more specifically.

5.2.2 Investigating the population: Hypothesis tests

A more indirect, but very popular method to find out some parameters or properties of the population is the hypothesis test. Hypothesis tests investigate whether our sample is chosen from a population with some specific properties.

For different questions different hypothesis tests are available. For example, if one wants to know if a sample comes from a population with a given mean (supposed that the population has a normal distribution), the so-called one-sample t-test should be used. Or if one wants to test if some data of two groups are coming from two populations where the means of the two populations differ (supposed that the two populations have normal distributions and they have the same variance), the two-sample t-test should be applied.

In the following paragraphs and subsections, we briefly review how the hypothesis tests work.

Null hypothesis. All hypothesis tests suppose a hypothetical population or populations with specific properties (e.g., the population is normally distributed), and with specific value parameter(s) (e.g., the mean of the population is 0, or the difference of the means of the two populations are 0 (i.e., the means are equal), etc.). This supposed situation is termed the **null hypothesis**. Then the hypothesis test investigates whether the actual sample could probably come from a population as described in the null hypothesis.

Sampling distribution. When the relation of the sample and the population was investigated we built up some intuition what could be some typical value when we sample from a specific population. Statistical analysis relies on this kind of information, but in a more systematic way. Let's see how this works in an example. Let's suppose that we want to see what typical mean values could be if the population is normally distributed, with a mean of 16 and SD of 5 (see the figure below), and the samples always include 20 cases. (For an excellent interactive online demonstration of sampling distribution see http://onlinestatbook.com/stat_sim/sampling_dist/ .) We measure a random sample, and calculate the mean. Then we repeat this sampling and mean calculation a few times. We might see again (as we've seen it before) that usually the sample mean is around the population mean, but occasionally we might get some farther value. The measured sample means can be shown in a distribution. This distribution is the **sampling**

distribution, where the name refers to the fact that the items of this distribution are some statistics of repeated samplings. If we repeat the sampling sufficiently many times (a few hundreds or even thousands of times), the distribution start to show some stable form: In this case, it is a normal distribution where the mean is the same as the mean of the population, and the SD is smaller than the SD of the population (see figure below). Now in this case, we might see precisely what sample mean values are more probable and what values are less probable.

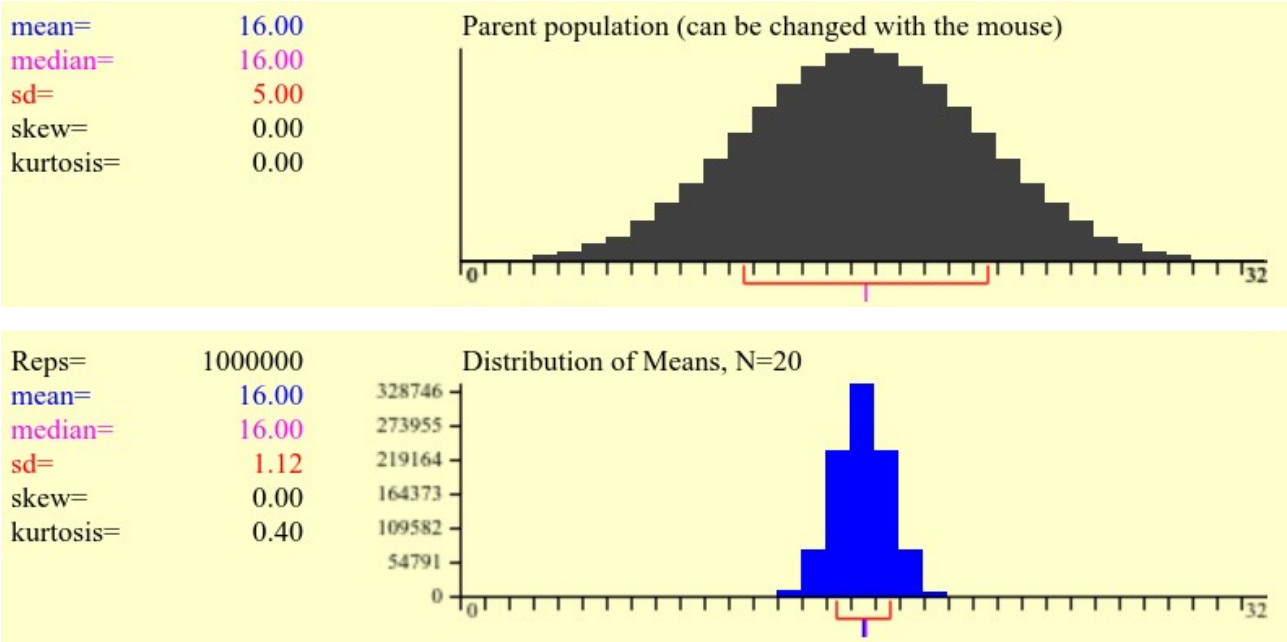


Figure A hypothetical population (top) and a related sampling distribution (bottom). Figure made with the online application by David M. Lane (http://onlinestatbook.com/stat_sim/sampling_dist/).

Extreme values in a sampling distribution and the p-value. Now let's suppose that we measure a new sample, and we want to see how extreme that sample is supposing that it comes from that population. One way to quantify how typical or how extreme a sample mean is when sampling from that population is to measure how large part of the sampling distribution is more extreme than the given value. For example, in the sampling distribution above value 18.5 is relatively rare and value 16.3 could be measured more frequently (knowing that the mean is 16 and the SD is 5 in the population). This extremity can be quantified. We could measure how frequently a value or more extreme value can be measured. For example, in the figure below, in the sampling distribution there is a value noted as "Observed data point". The green area denotes the probability of getting that value or more extreme values. Now the chance of getting that value or more extreme values is the ratio of the green area compared to the whole area under the sampling distribution function. In other words, this ratio index will tell us what percentage of the whole sampling distribution area is green. This extremity index is the **p-value**. The p-value can be between 0 and 1 (i.e., between 0% and 100%). The smaller the value is, the less probable that the actual sample is from the hypothesized population. Thus, the p-value will quantify the extremity of a specific sample in that sampling distribution. In other words, the p-value specifies that if the null hypothesis is true, how extreme our sample would be in possible samples from that population.

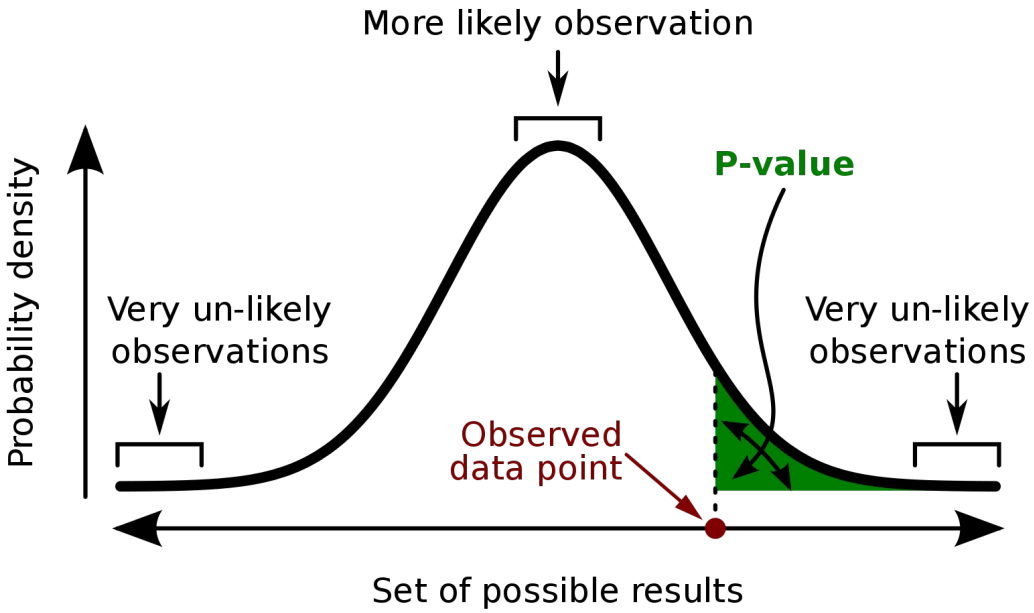


Figure Sampling distribution with an observed data point and with the probability of finding such or more extreme value (green area) in that sampling distribution (source)

Testing other null hypotheses. The example above used the mean to create the sampling distribution of the mean to find the extremity of a sample mean in that sampling distribution. This example can be generalized to other cases, forming the general method hypothesis tests rely on. So one can form a population (or set of populations), and instead of calculating the mean, any other formula could be used. Then the sampling distribution can be specified. Finally, if one wants to use that method, the same formula should be used for the measured sample, and the extremity of that value in the sampling distribution should be calculated, which extremity is the p-value. Note that statisticians already created a lot of hypothesis tests that test various null hypotheses, and the statistical software can run many of those tests and return the appropriate p-values.

Rejecting the null hypothesis. The p-value could be used directly, and for example, one could say that the sample I've just measured would be quite extreme (my p-value is 7%) to come from the null hypothesis (i.e., from the hypothetical population), so probably the sample is not coming from that hypothetical population and the null hypothesis is not true. However, it might be debated what value could be considered as extreme or non-extreme. To overcome this ambiguity, researchers agreed upon that p-values that are smaller than 5% will be considered as extreme enough to support the idea that the null hypothesis is not true. (Note that 5% was chosen not because it is the best available value for this purpose, but because it was in a reasonable range, and was chosen so different studies could be comparable.) So based on this convention, if a hypothesis test returns a p-value that is smaller than 0.05, it is concluded that the null hypothesis is not true (or in the terms of statistics, **the null hypothesis is rejected**, and in other terms, the result is significant).

Type-I error. The null hypothesis rejection introduces a new problem because in this case the null hypothesis is considered to be false, even if there is a 5% chance that the null hypothesis is in fact true (i.e., although in a few cases, but the sample could be as extreme as well). In other words, in

5% of the time the null hypothesis will be rejected, even when the null hypothesis is in fact true. This error is called **type-I error.**

Threshold value for rejection. Could we decrease the chance of type-I error with lower threshold level, such as 0.01, or 0.001. Why don't scientists use a lower threshold? Actually, this would introduce two other problems. (1) The sampling distribution might include arbitrarily large or small values. Remember, for example, a normal population can produce an arbitrarily extreme sample as we have seen above. Similarly, in a sample distribution extreme values could take place, and therefore, if one wants to be 100% sure that the sample to be tested does not come from the null hypothesis population, then no sample statistics could be considered as impossible. In other words, if one wants to be 100% sure, then one could never reject the null hypothesis, which makes the whole procedure pointless. (2) Lower level extreme value threshold would decrease the so-called type-II error, which error is not explained in the present text, but which problem we also want to avoid. Thus, one has to balance between incorrect rejection of the null hypothesis (i.e., type-I error) and other problems, and in practice the 0.05 threshold level seemed to be a relatively reasonable compromise.

Not rejecting the null hypothesis. While with a smaller than 5% p-value the null hypothesis is considered to be false, you might think that with higher than 5% value the null hypothesis is true, because samples like our sample could come quite frequently from that population. Actually, this is not true, and if the p-value is larger than 5%, the null hypothesis is not considered to be true. (Unless an indirect trick is used with power analysis which we don't discuss in the present text). Instead, with larger than 0.05 p-value it is reasoned that it is not known whether the null hypothesis is true or not. Here is why. Imagine an interval scale where the values can be between 1 and 5, and we want to test whether the population mean value is 2. With a hypothesis test a p-value is calculated, and let's say that it is 96%. One could argue that this hypothesis test result means that it is very typical that a population as described in the null hypothesis generates a sample like ours, so our null hypothesis is probably true (i.e., the population mean is 2). However, this reasoning in itself is incorrect. The main reason why this is incorrect is that a different null hypothesis that is very similar to the previous null hypothesis (e.g. a population with a mean of 2.001 instead of a population with a mean of 2) will give a very similar p-value. So by the same logic we have applied for the test value 2, one should say that this new null hypothesis of 2.001 value is also true. The problem here is that these two conclusions lead to a contradiction because both null hypothesis cannot be true at the same time. So for these reasons, when our p-value is larger than 5%, the only strictly appropriate conclusion is that we cannot say that the null hypothesis is false, but it is either true or some similar null hypothesis is true (and at the same time we don't know how different that other hypothesis is from the original null hypothesis).

Summary. To summarize so far, in a hypothesis test a null hypothesis is specified, and then the test can calculate the chance that the actual measured sample could come from that hypothetical situation, where this chance is termed the p-value. If that p-value is small (usually if it is smaller than 5%), then the null hypothesis is considered to be false. On the other hand, if the p-value is larger than the critical level (i.e., usually larger than 5%), then no conclusion is drawn because either the null hypothesis test is true, or some similar alternative hypothesis is true.

Main issues of the hypothesis tests. Although this summary sounds reasonable, the whole procedure introduces a few problems. First, type-I error was introduced, so sometimes (on average 5% of the time) we will reject the null hypothesis even if it is true. Note that this is another example,

why you never can trust inferential statistic 100%. Second, when the null hypothesis is not rejected, it cannot be accepted at the same time, because if no other details are known, then we couldn't know if the null hypothesis is true or some other, similar null hypothesis is true. Relatedly, although we didn't discuss it here, sometimes when the null hypothesis is not true, the test does not reject it. Overall, while the hypothesis test is a popular tool to analyze the data, there might be problems with its results (and because of those problems many researchers misunderstand its meaning and misinterprets the results), so one has to be cautious when using it.

5.2.2.1 Choosing the appropriate hypothesis test

For the various questions and other circumstances there are different hypothesis tests, and it is critical that the appropriate hypothesis test should be chosen. CogStat For most statistical software the user have to know which specific test should be used. In contrast, CogStat chooses the hypothesis tests automatically. It is useful to understand the main decision viewpoints how tests are chosen, but CogStat users don't have to know all specific rules and hypothesis tests, because CogStat takes care of that.

Research questions. One key factor that the choice of the appropriate hypothesis test depends on is what is the research question. Do we want to compare groups? Do we want to investigate the relation of variables? Do we want to see the mean of a single variable? For different questions there are different hypothesis tests. Note, however, that there could be some questions for which no hypothesis tests have been created yet. This should be considered when one creates the research plan: The measured data should be in a form that can be investigated with the available statistical methods (or more rarely, one should be able to create new tests that fits the research question).

Measurement level. Another key factor is the measurement level of the variables. For example, comparing central tendency of two groups requires two-sample t-test for interval dependent variables (but see generally the assumptions of the tests below, and specifically the assumptions for the two-sample t-test here), but it requires Mann-Whitney test for ordinal variables. On a more technical level, the questions of these different tests are also different. While in the previous example the central tendency was investigated, in fact, for interval variables this central tendency is the mean, but for ordinal variables it is the median.

Assumptions Some of the hypothesis tests work correctly only if some preconditions apply. In some way, the research question and the measurement levels are also such preconditions, but there are some specific properties of the data that should be present, otherwise the tests will not work correctly. For example, several hypothesis tests assume that the variables are normally distributed. It means that the test will work correctly and will give correct p-value only if the appropriate variables are normal. If the appropriate variables are not normal, then the returned p-value will be incorrect and biased – if the assumptions do not apply, the hypothesis test should not be used.

Assumptions of the specific tests. It is a specific property of all hypothesis tests what assumptions should be applied. Some tests do not need any assumptions about the data, for example, the Kruskal-Wallis test does not require any assumptions. Some others need at least one assumption. If there are several assumptions, then all of them should be fulfilled. In other words, if

any of the assumptions are violated, then the main hypothesis test cannot be run. Overall, all hypothesis tests specify what assumptions should be ensured to make the test work correctly.

Checking the assumptions. How can these assumptions be checked? There are several methods to check or tell whether the appropriate assumptions apply. Probably the most common solution is to run additional hypothesis tests that check the property of the assumption. For example, there are hypothesis tests for checking if a variable is normally distributed, such as the Shapiro-Wilk test. Usually the null hypotheses of those test say that the assumption is not violated. So if the tests for the assumptions are significant (i.e., the p value is smaller than 0.05 (5%)), then the assumption is violated, and the hypothesis test for which the assumptions were checked cannot be run.

When assumptions are violated. What happens if some of the assumptions are violated? It is clear that the hypothesis test for which the assumptions were checked cannot be used, but what shall we use instead? In those cases, usually there are alternative hypothesis tests that can be used, and which tests do not require the assumptions the original test required. For example, if the assumptions of the one-way ANOVA (for comparing the mean of more than two groups with interval dependent data) are violated, Kruskal-Wallis test will be run. Now if some hypothesis tests do not require strict assumptions, why don't we use them in the first place? Among the hypothesis tests that investigate similar properties, the ones with more assumptions are usually more sensitive, i.e., they can reveal an existing effect more probably, than the ones without assumptions. The main reason why the tests with more assumptions are more sensitive is because those suppositions make the investigated situation more specific and the tests might rely on more information compared to the tests without assumptions. Overall, we want to use the more sensitive tests first, supposing that the assumptions apply, otherwise we might switch to the less sensitive "backup" test.

Based on the factors listed so far we might understand now how the hypothesis tests are selected. Still, the story is not over because beyond knowing the general considerations, one should know the various tests in various conditions. Usually, most textbooks introduce the most common tests and their preconditions. Also there are descriptions that summarize the possibilities, for example see https://statkat.com/stattest_overview.php, http://www.socr.ucla.edu/htmls/SOCR_ChoiceOfStatisticalTest.html or https://stats.idre.ucla.edu/other/mult-pkg/whatstat/.

CogStat Note again that CogStat chooses the hypothesis tests automatically. It is important that you should understand the main decision viewpoints how tests are chosen, but it is not necessary to know all the specific rules and hypothesis tests, because CogStat will handle that.

6 Analyses result sections in CogStat

CogStat

In CogStat, most es results include three sections: the raw data (together with the number of the available cases), the sample properties, and the population properties. Here we summarize the main components and sections of the results of an analysis.

Exercise Import the data found on the Exercise data sheet of the supplementary file, and choose the Analysis > Compare groups task. Set the After variable as the dependent variable and the Region variable as the grouping variable, and run the analysis. Check the result section in your results as described below.

Number of available cases. Whenever we measure a sample, there could be occasions where the data of some cases are not available, so those data are missing. There could be many reasons why this could happen. For example, the participant denied replying some of the questions, some part of the log file was accidentally lost, some clearly incorrect value was recorded which should be ignored, and so on.

CogStat displays the number of valid cases and the number of missing cases. Valid cases are the cases whose values can be used in an analysis, and CogStat will use only those valid cases.

Raw data graphically. Then, the raw data are displayed graphically, without displaying any further sample or population properties. This might be useful to detect some important features of the sample that could not be observed based on the calculated sample or population properties, such as outlier values or bimodal distribution (i.e., a distribution with two peaks, technically, with two modes), etc. (find more information here why looking at the raw data is useful). Also, looking at the raw data we might have an impression what results we could get with the current data.

Sample properties. In the sample properties part, first, a numerical summary of the sample is shown, describing key properties of the sample (descriptive statistics), such as the mean for an interval variable. Then, a graphical presentation of some key features of the sample is displayed, for example, the box plot or histogram of an interval variable. Usually, in those charts the raw data are also visible. Note again that these details are related to the sample and not to the population.

Population properties. The last section of the output includes the properties of the population. These values are always estimations and results of indirect reasoning, so one never can be sure whether these values are precise or reflect the real population values faithfully. Usually, these population properties show some point and interval estimations of the population parameters, and show some hypothesis tests about various properties of the population.

Advanced CogStat In most software packages, the results are not grouped or labeled as sample and population properties. Still, it is important to understand whether a property relates to the sample or the population because they should be interpreted differently. Sample data are the hard fact in a sense that they are the values we have measured and we rely on in our analyses. Note also that sample data are a somewhat distorted reflection of the population, which causes some uncertainty in their interpretation. Consequently, population properties (that rely on the sample data) are just speculations and although most of the time they might be correct estimations and tests, they never can be 100% percent correct.

6.1 Measurement levels in CogStat charts

The chart axes reflect the measurement levels. Interval variables are displayed as solid lines, ordinal variables are displayed as dashed lines, and nominal variables are displayed as dotted lines. See more details in the CogStat documentation.

7 Exploring a single variable

The simplest analysis is to measure a single variable (property) in a single group (which sample is part of a population), and to investigate various mathematical and statistical properties of that variable.

CogStat In CogStat, use the 'Analysis > Explore variable' menu to perform such an analysis. In the dialog, choose the variable you want to investigate, then click OK. (See the general structure of the output in CogStat here. See the details of the displayed results in the CogStat documentation.)

7.1 Raw data

Raw data are the valid (i.e., non-missing) data that we want to analyze. All following analyses rely on these values.

The meaning and interpretation of those data depend on the measurement level of those data. Therefore, depending on the measurement level, different charts will be used to display those data.

7.1.1 Interval raw data

Exercise In the Analysis > Explore variable menu set After to be a selected variable, and run the analysis. Check the results of your analysis according to the description below. After looking at your raw data try to estimate your sample and population properties. (Note that the example illustrations may include other data than the Exercise data.)

For an interval variable, the relevant information is the values of the numbers. (This sounds trivial, but for an ordinal variable, the relevant information is the order of the numbers and not the value of the numbers. See Ordinal raw data. Also, in nominal data, the values are only important in a sense whether they differ or not. See Nominal raw data.)

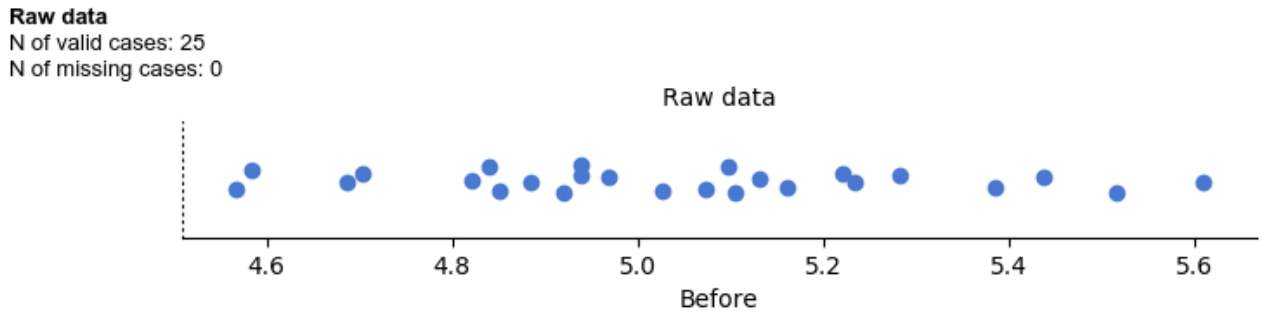


Figure Raw data of an interval variable

CogStat The raw data are displayed in a dot plot. In CogStat, the scale is a horizontal number line, and dots are displayed to denote the data.

When there are values that are close to each other, sometimes it is hard to see these overlapping signs, so sometimes some random jitter is added, to make these overlapping data more visible.

When there are multiple cases with the same values, the size of the sign is proportional to the number of cases. (There are data analysis software packages that do not denote if there are multiple cases with the same values, which charts cannot show the data accurately.)

In most cases, we can observe that there is some central region where most of the cases are placed, while moving away from this central part the data become less frequent. In fact, many times, the distribution follows a very specific pattern, the so-called normal distribution (see more information about normal distribution here).

7.1.2 Ordinal raw data

Exercise In the Analysis > Explore variable menu set Preference_night to be a selected variable, and run the analysis. Check the results of your analysis according to the description below. After looking at your raw data try to estimate your sample and population properties.

For an ordinal variable the relevant information is the order of the values. Recall that one cannot tell what the distance is between two neighboring values (or more generally between any two values) (see Measurement levels).

CogStat To display the relevant order information, CogStat displays the order of the values and not the values themselves. The chart is a dot plot, like in the case of interval variables, but there are critical differences compared to the interval dot plot. First, the x axis shows two numbers: first, the order of the data, then in parentheses the original value. Note, that the chart is created based on the order and not on the value. This might look unusual for the first time, but in fact, it is the order of those values that will be considered in the following analyses, and it is more appropriate to display the information the variable includes and consequently the analyses rely on. Second, the line of the axis is dashed. It reflects that the axis cannot be considered as a continuous line (as used for interval variables), but there are unknown distances between neighboring values. Dashed lines can remind you that this variable is not an interval scale.

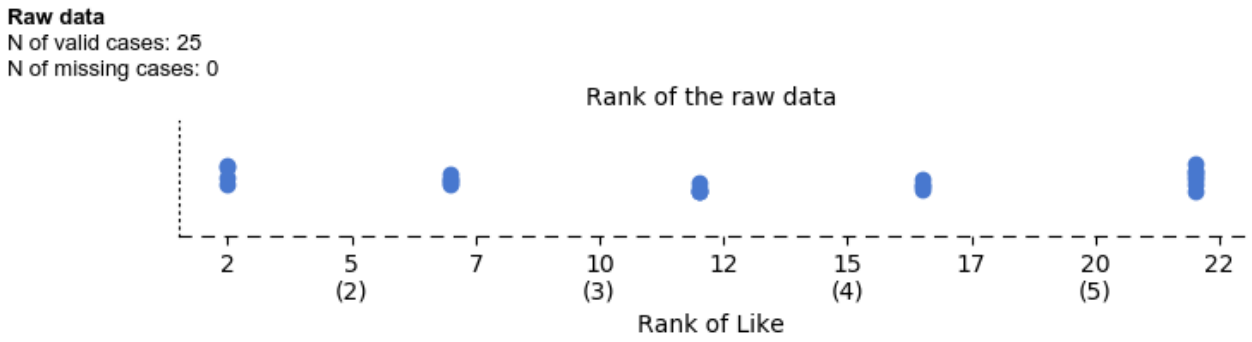


Figure Raw data of an ordinal variable

7.1.3 Nominal raw data

Exercise In the Analysis > Explore variable menu set Region to be a selected variable, and run the analysis. Check the results of your analysis according to the description below. After looking at your raw data try to estimate your sample and population properties.

For a nominal variable the only relevant information is whether a case can be described by a property or not, and the values only reflect this property. If the values of the nominal variables are

expressed as numbers (and not as letters, words, or other non-numerical symbols), then the order or the distances of the values cannot be used. For this reason, the only raw data information is the number of cases with the specific values. CogStat displays a histogram of the variable, where the frequency of all values is displayed. Similarly to the special status of ordinal data which were denoted by dashed axis in the charts, when nominal data are displayed on a chart in CogStat, the appropriate axes are dotted. Dotted axes hint that the axis is not really a continuous dimension, or not even a dimension where things could be ordered.

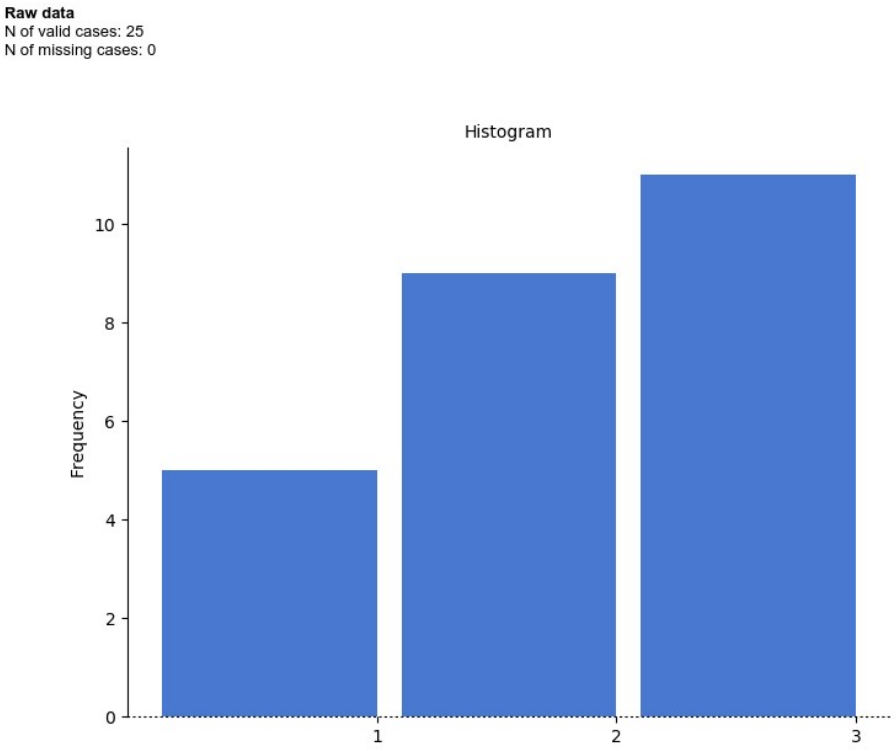


Figure Raw data of a nominal variable

7.2 Sample properties for interval data

7.2.1 Frequencies table

The frequencies table is a very simple summary of the various values of the sample. First, it lists all the **values** that were measured in that sample (value column) in increasing order. Then, it presents the **frequency** (frequency column), which is the number of those values in that sample, i.e., how many cases had that specific value. The **relative frequency** (relative frequency column) is the same information as the frequency, but not as a count value, but as the percentage of that frequency in the whole sample. For example, if the sample included 20 valid cases, and the value 3 was measured 4 times, the relative frequency of the value 3 is 4/20=0.2 or 20%.

The **cumulative frequency** includes not only the frequency of that value, but the sum (therefore the name, cumulative) of the frequencies of the current and all the smaller values. This also means that for the largest value, the cumulative frequency should be the sample size because all other values are smaller, and the sum of the frequency of all values should be the sample size (or at least the number of valid cases if there are missing cases in that sample). The **cumulative relative**

frequency is again the same information as the cumulative frequency, but it is expressed as the percentage of that value of the sample size. For the largest value, the cumulative relative frequency should be 1.0 or 100%.

Frequencies

Value	Freq	Rel freq	Cum freq	Cum rel freq
4.566828	1	4.0%	1	4.0%
4.584376	1	4.0%	2	8.0%
4.687729	1	4.0%	3	12.0%
4.703795	1	4.0%	4	16.0%
4.820753	1	4.0%	5	20.0%
4.839246	1	4.0%	6	24.0%
4.850750	1	4.0%	7	28.0%
4.884206	1	4.0%	8	32.0%
4.920813	1	4.0%	9	36.0%
4.938754	1	4.0%	10	40.0%
4.939499	1	4.0%	11	44.0%
4.967994	1	4.0%	12	48.0%
5.027725	1	4.0%	13	52.0%
5.072816	1	4.0%	14	56.0%
5.097055	1	4.0%	15	60.0%
5.104881	1	4.0%	16	64.0%
5.131748	1	4.0%	17	68.0%
5.161934	1	4.0%	18	72.0%
5.220715	1	4.0%	19	76.0%
5.234405	1	4.0%	20	80.0%
5.283282	1	4.0%	21	84.0%
5.385876	1	4.0%	22	88.0%
5.438096	1	4.0%	23	92.0%
5.516184	1	4.0%	24	96.0%
5.610113	1	4.0%	25	100.0%

Figure Frequencies table of an interval variable

7.2.2 Descriptives

See an earlier explanation why sample statistics is useful to describe some main properties of the sample in a concise way instead of looking at the raw data.

In the following subsections we describe all the relevant descriptives used for the interval variable of the sample.

Descriptives for the variable

	Before
Mean	5.03958297375954167
Standard deviation	0.27190562561982218
Skewness	0.24419377130974071
Kurtosis	-0.43118681946248305
Range	1.04328503861858035
Maximum	5.61011300071542962
Upper quartile	5.22071522276081001
Median	5.02772537468387970
Lower quartile	4.85074989045562965
Minimum	4.56682796209684927

Figure Descriptive values of an interval variable

7.2.2.1 Central tendency

For an interval variable there are several summary properties that can be important. Central tendency is one of the most important properties (and one of the most frequently used properties). It shows where the distribution is located on the scale. For a distribution that is similar to a normal

distribution it shows where the middle of the distribution is located. One simple central tendency index is the median. The **median** is the value of the case for which the same number of cases is larger and smaller. In other words, if the sample values were ordered, the median is the middle value. Another better-known central tendency index is the mean. The **mean** is calculated as the sum of all values divided by the number of cases.

Many times the mean and the median are close to each other. However, there are some reasons why they may differ. First, the median can only be a number that is a measured value in the sample, while the mean can be other values, too. For example, in a small sample including only 3 values, 4, 5 and 9, the median is 5 (a value that was also measured), while the mean is 6 (which is not among the measured values.) (Note that when the sample size is even, e.g., there are 4 cases, and there is no middle case, according to some median calculation methods, the median is the mean of the two middle cases.) Second, in some distributions (for example, in the frequently observed normal distribution), extreme values (i.e., values that are far from the median or the mean) change the mean stronger than the median, so median is less sensitive to those extreme values. (You can imagine that when any value is moved to the more extreme part of the distribution, this change will modify the mean, while the median will not be changed.) That is one of the reasons why the median is preferred in some central tendency calculation, as the median is "less noisy".

Advanced Some texts mention and some researchers argue that median should be calculated only for ordinal variables, but not for interval variables. This is not true: For any measurement type variable you can also calculate the statistics of the lower measurement level statistics, too. For example, you can calculate the median for an interval variable. Calculating the statistics of a lower measurement level has advantages and disadvantages: on one hand using only the information of a lower measurement level means losing information. For example, median do not utilize the value information, but only the order information of the variable. On the other hand, lower measurement level statistics might be beneficial. For example, the median is less sensitive to the extreme values of a distribution that the mean, so the median might be less noisy than the mean. As another example, Spearman correlation does not require the assumption that are required by the Pearson correlation.

7.2.2.2 Spread

The spread can tell the extent of the data on the scale, or in other words, how widely one can find the data on that scale.

A simple way to measure the spread is to calculate the range. **Range** is the difference of the smallest and largest value in a sample. In other words, it is the whole range where the data takes place. A more complicated, but frequently used spread index is the standard deviation. **Standard deviation** (often abbreviated as SD) shows how "wide" the distribution is on the scale. SD is calculated as the (a) square root of (b) the mean of (c) square distance from the mean (see figure below for the formula). This sounds a bit complicated, but the formula is not so scary. For the SD the formula calculates (c) the distance of all cases from the mean, then those differences will be squared, then (b) those squared values will be averaged, and finally, (a) this mean will be square rooted. SD is a kind of mean distance from the mean, but mathematically it is not the mean of the distance, but a bit less intuitive calculation, as explained above. Still, it is an index that is used very frequently.

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2},$$

Figure Formula of the sample standard deviation (source)

7.2.2.3 Skewness and kurtosis

Skewness is an index which tells how asymmetric a variable is. If the value is zero, then the distribution is symmetric. If the value is positive, then the distribution has a longer tail on the right side, and if the value is negative than the distribution has a longer tail on the left side.

Kurtosis is an index that measures the “tailedness” of the distribution: it shows whether the tails are relatively thick with smaller part in the middle or whether the tails are thin with larger middle part of the distribution. For a normal distribution the kurtosis is 3. If the index is smaller than 3, then the distribution has less outlier or extreme values and includes more items around the mean.

Skewness and kurtosis are rarely investigated for themselves, but they might reflect whether the distribution is normal (skewness is 0 and kurtosis is 3), so they are occasionally used as a measure of normality.

Advanced Many times only by looking at the distribution it is hard to tell whether a distribution has thinner tails or whether the standard deviation is small. The skewness and the SD are two distinct properties, but by looking at the distribution it might be hard to tell how strongly those two properties contributed to the shape of the distribution. To illustrate this point, have a look at the figure below, and try to estimate whether the blue or the orange distribution has larger kurtosis and which one has larger SD. Read on after you have your guess. Probably the easier part is the kurtosis: because the blue one has thinner tail, it might have larger kurtosis, and indeed, the blue distribution has a 9 kurtosis, while the orange has 3. The more surprising part (or maybe it is not so surprising anymore, because you know that we are trying to prove a surprising point) is that while you might guess that the orange one has larger standard deviation because it is wider than the blue one (at least in the middle part of the distribution), in fact its 0.8 SD is smaller than the 0.9 value of the blue distribution.

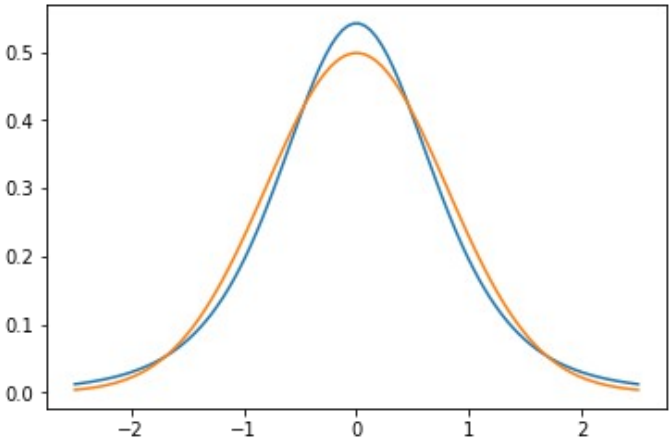


Figure Two distributions with some hard to visually estimate parameters

7.2.2.4 Minimum, maximum and percentiles

Minimum is the smallest value and maximum is the largest values in a dataset.

Percentiles are specific values below which a specific percentage of the data falls. For example, 30th percentile is the value below which the 30% of the values fall. In other words, imagine that the values of the data are ordered, and the value is chosen where x% of the data fall below and 100-x % of the data fall above.

There are some special percentiles. One of such percentiles is the median. Median is in fact the 50th percentile because half of the data are below that value and half of them are above it. Other special percentiles are the lower and upper quartiles that are the 25th and 75th percentiles. You might imagine that the two quartiles and the median splits the data into four subsets with equal number of cases.

7.2.2.5 Histogram and box plot

CogStat In CogStat, in the chart presenting the sample properties, different sample properties are displayed in a single chart. First, the raw data can be seen. This is the same information that could be seen in the Raw data section, although in a bit different format. Instead of using dots, the raw values are small lines on the lower axes of the upper part of the chart. Second, the histogram of the data is displayed. Third, the box plot is displayed in the lower part of the chart. In a **box plot**, the middle of the inner box is the median, the width of the inner box displays the interquartile range, which is the range between the lower and upper quartiles, while the whiskers display the range. This means that the box plot shows the four ranges of the data where all ranges include the same amount of cases.

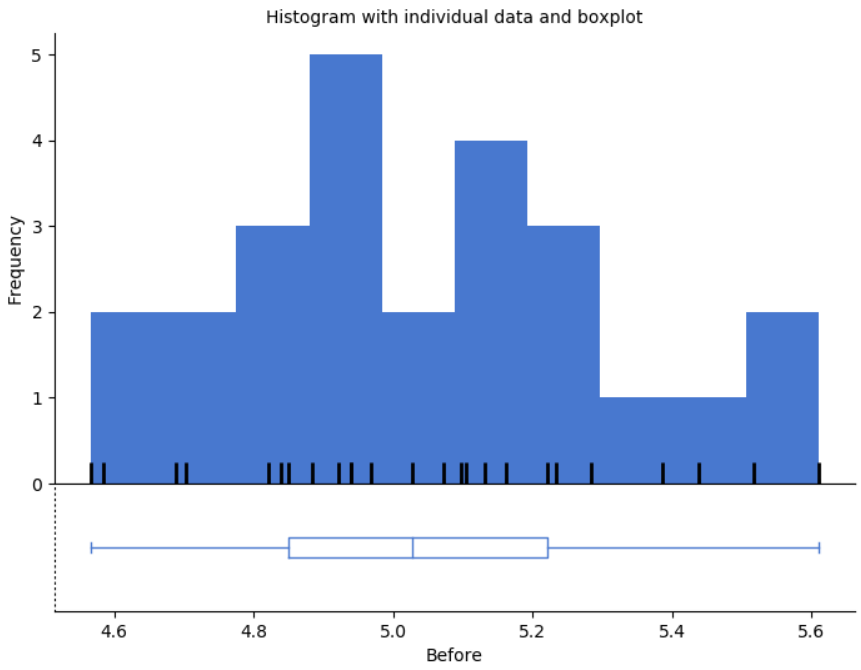


Figure Histogram and box plot of an interval variable

7.3 Sample properties for ordinal data

For the frequency table the same information is included as for an interval variable.

For an ordinal variable less descriptives are calculated than for an interval variable. For an ordinal variable, only the median is used as a central tendency: Because for an ordinal variable the order of the values can be used, it makes sense to pick the middle value, i.e., the median. On the other hand, the mean cannot be calculated because the distance between the values of an ordinal scale is meaningless, so it would not make sense to sum those values (i.e., the distance from the zero point).

Descriptives for the variable

	Like
Maximum	5.0
Upper quartile	5.0
Median	3.0
Lower quartile	2.0
Minimum	1.0

Figure Descriptives for ordinal data

To display the spread of the data, minimum and maximum is displayed, and also upper and lower quartile, similar to the interval variables. However, the range (the difference of the minimum and maximum) and the interquartile range (the difference of the upper and lower quartile) is not shown because in ordinal variables it does not make sense to subtract the values since the distance of the values is not defined. (For similar reason, skewness and kurtosis cannot be calculated either.) Advanced Note that while some descriptions suggest that the spread of an ordinal data can be measured with the range of it (i.e., the difference of the smallest and the largest values), for the reason mentioned above it doesn't make sense to calculate the range (it includes the difference of two values, which subtraction is not supported for ordinal data).

CogStat In CogStat, for an ordinal variable, the sample properties chart is similar to what can be seen in the sample properties of the interval data, but the ordinal data chart relies on the order information. Consequently, the chart displays the order of the values and the relevant chart axis is dashed.

When most cases differ in their values and consequently in their order (i.e., there are no multiple cases with the same value, and therefore, with the same order), the histogram is uniform. In that case, for the same reason, the box plot will show four approximately equal length quartiles.

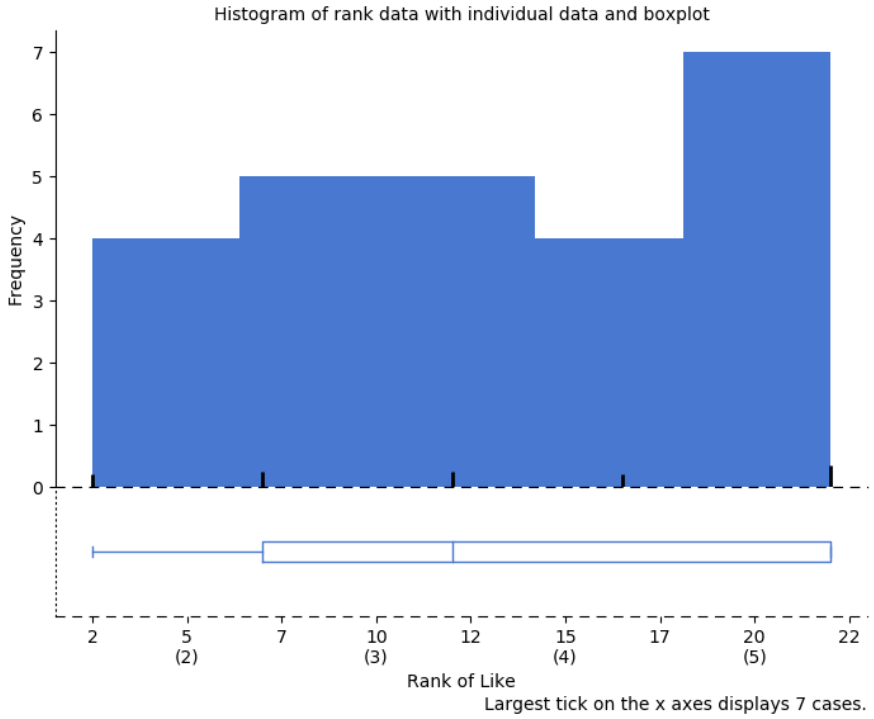


Figure Histogram and box plot of an ordinal variable

7.4 Sample properties for nominal data

First, the frequencies of the values are displayed. However, unlike in interval and ordinal data, cumulative frequencies are not displayed, because nominal data cannot be ordered, therefore, cumulative frequencies across increasing values would be meaningless.

For a nominal variable, a lot of summary statistics that has been shown for interval and ordinal variables would be meaningless because in a nominal variable, the different values are entirely unrelated of each other, and they cannot be summarized according to their order or value, since their relation cannot be specified. For example, for a nominal variable it does not make sense to talk about the central tendency, or spread because the values of the variable cannot form a series to create such descriptives.

CogStat In CogStat, at the moment, it is only the frequencies table that is shown for the sample properties.

Advanced Sometimes it is stated that the mode (the most frequent value) can be calculated as a central tendency for a nominal variable. This is misleading because the mode indeed could be calculated for a nominal variable, but it cannot be a central tendency in a sense that other values are around it, and in a sense that the mode would be the center of the distribution. The main reason why the mode cannot be a central tendency index is that generally a nominal variable cannot have a central tendency. The misunderstanding could come from the fact that in several distributions of interval variables, such as the normal distribution, the mode is usually close to the mean and the median, and for such an interval variable it is really a central tendency. However, nominal variables do not have central tendency, and the mode cannot be a central tendency in that sense either.

7.5 Population properties for interval data

As we have mentioned, one of the main issue in data analysis is that we can only see a sample, which is a small part of the population, but we want to know the properties of the population that we cannot observe directly.

7.5.1 Normal distribution

One property that is evaluated is whether the population distribution our sample data come from has a normal distribution. Most of the time, researchers are not interested whether the population of the sample is normal or not in itself, but they want to know whether some other more sensitive analyses can be performed which analyses can be used only if the data are coming from a normally distributed population. Still, in some special cases, researchers might be interested in whether a variable is normally distributed or not because that could be informative about some specific theoretical question.

In CogStat, the Shapiro-Wilk hypothesis test is run as a hypothesis test checking whether a distribution is a normal distribution. In a normality test, the null-hypothesis is that the distribution is normal, so a significant result means that the data is probably non-normal. There are many hypothesis tests that could be used to test the normality of a distribution, but according to methodological studies the Shapiro-Wilk test is among the best ones.

In a chart, the distribution of the data and the raw data are displayed, like in the sample properties result section. The extra part is the normal distribution line, which has the same mean and standard deviation as the sample. So if the normal distribution line goes together with the columns of the histogram, then the sample may come from a normally distributed population.

Normality
Shapiro-Wilk normality test in variable Before: $W = 0.981$, $p = 0.896$

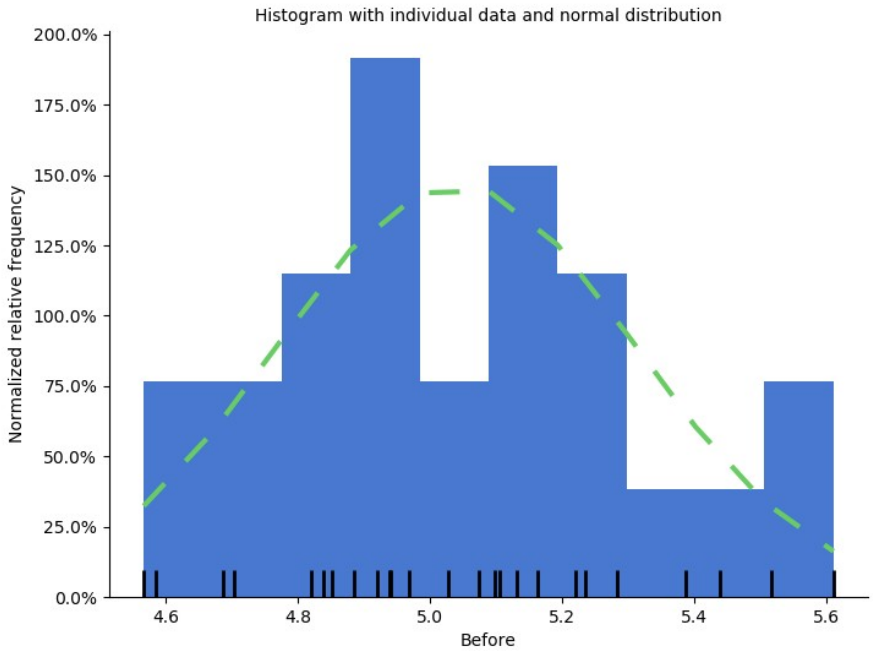


Figure Histogram of an interval variable with normal distribution

Another way to check visually the normality of the data is to see the quantile-quantile plot or Q-Q plot. The Q-Q plot is a scatter plot. Practically, if the dots can be seen near the diagonal line, and no systematic deviation can be seen, then the sample is probably normally distributed. The figure relies on quantiles, which can be considered as what percentage of the data can be found below the actual value. The x axis shows the quantiles of the normal distribution in z-scores, and the y axis shows the quantiles of the data. (Z-score or standard score is a transformation of some data. For a set of values all values are transformed with the formula (x-mean) / standard_deviation, where x is the value, and mean and standard deviation is the descriptives of the set. Conceptually, z-score will tell how many standard deviations the value (raw score) is above or below the mean. E.g., in a sample with a mean of 5 and SD of 2, a raw score of 1 will be transformed to a -2 z-score, which z-score means that the original value is 2 standard deviations below the mean.) In the scatter plot all quantiles for the data are calculated (e.g., what percentage of the sample is included if only the smallest value is considered, what percentage for the two smallest values, etc.), and the same quantiles for the normal distributions are calculated (what z-score that distribution has for those quantiles), then the normal distribution z-scores and the sample values are displayed as a scatter plot. With this procedure if the sample looks similar to the normal distribution then the dots of the scatter plot should be around the diagonal.

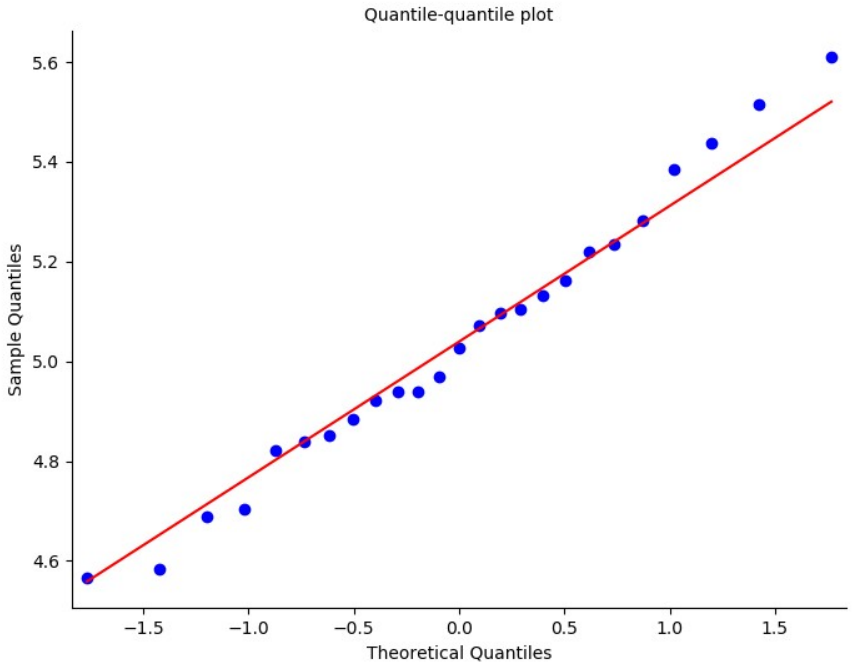


Figure Normal quantile-quantile plot

Many times, the graphs and the hypothesis tests are in line: either both of them suggest that the distribution is normal (in the Q-Q plot the dots are around the diagonal line and the hypothesis test is not significant) or both of them suggest that the distribution is not normal (in the Q-Q plot the dots systematically deviate from the diagonal, and the hypothesis test is significant). However, in some cases, while the graphs suggest a non-normal distribution, the hypothesis test is not significant. This usually happens, when the sample is small. You can think about this as although the graphs suggest non-normality, the sample is small enough to not give a convincing amount of evidence for non-normality because in small samples anything could happen. The reversed pattern can also be observed: The Q-Q plot suggest that the distribution is normal, although the normality

hypothesis test is significant. This might happen when the sample is large, and the deviation from normality is not so strong, but the hypothesis test is more likely to be significant for relatively small deviation when the sample is large.

7.5.2 Population estimations

For interval data, the population estimation of the mean and standard deviation is calculated. For the mean, it is the same value as the sample mean. For the SD, the population estimation is a bit larger than the sample SD (see this page for more information about the differing formula).

For the mean, CogStat also calculates the 95% confidence interval. This means that most probably the population mean is in that range. However, as always, this is just an estimation, and there is still a small chance that the population mean is somewhere outside of this range. The mean estimation with the 95% CI is also displayed graphically. (Note that related information is provided by the appropriate hypothesis test.)

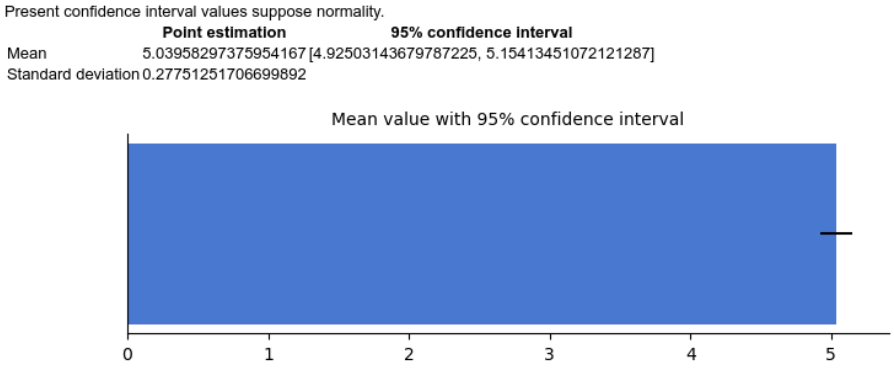


Figure Population estimations of an interval variable

7.5.3 Hypothesis test for the mean

One of the most frequent statistics that researchers are interested in is the mean of the variables. More formally, they want to know the mean of the population the data came from. This has already been investigated with the 95% CI of the mean, but there is another way to investigate it: To use a hypothesis test.

In a hypothesis test, where the mean of the population is tested, the null hypothesis is that the population has a specific value of the mean, e.g., 0, and a significant result means that this is not the case. CogStat The specific value can be set before running the analysis, for example, in CogStat, in the 'Explore variable' dialog you can change the 'Central tendency test value' from 0 to any other value.

Most usually, either the so-called one-sample t-test is used (if the variable is normally distributed) or the Wilcoxon signed-rank test (if the variable is non-normal). CogStat CogStat automatically tests whether the data is normally distributed or not, and depending on the result, it will choose the appropriate test to check the mean value (i.e., either the one-sample t-test or the Wilcoxon signed-rank test).

Hypothesis test: Testing if mean deviates from the value 0.
Interval variable. >> Choosing one-sample t-test or Wilcoxon signed-rank test depending on the assumption.
Checking for normality.
Shapiro-Wilk normality test in variable Before: $W = 0.981$, $p = 0.896$
Normality is not violated. >> Running one-sample t-test.
One sample t-test against 0: $t(24) = 90.8$, $p < 0.001$

Figure Hypothesis test for an interval variable

Advanced One issue with the normality hypothesis test is that for small sample the test might lack power (i.e., in a small sample it is impossible to tell if the distribution is not normal). For this reason, the normality check might be imprecise. Could we trust the results in this case? Several researchers and statisticians suggest that the hypothesis test shouldn't be used for normality assumption checks. Still, as far as we know, the alternative methods are even riskier, so the hypothesis test might be the least worst solution for checking the normality of a variable.

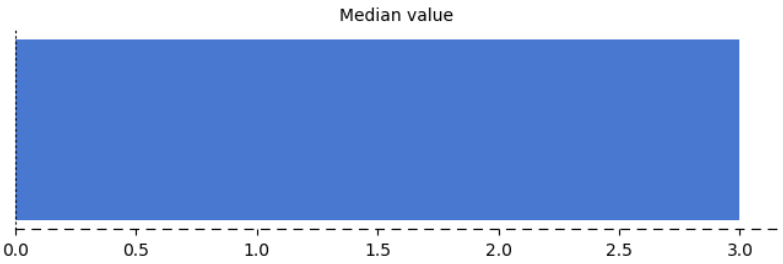
Advanced The confidence interval of the mean and the hypothesis tests for the mean investigate the same thing: The mean of the population. The CI includes the probable population means, and the hypothesis test tells whether the test value is an unlikely population mean (i.e., when the result is significant). Importantly, they not only investigate the same thing, but their results are closely related and are consistent. The results of the two methods are in line: When the CI includes the test value of the hypothesis test, the hypothesis test is not significant, and when the test value is outside of the CI, the hypothesis test is significant.

7.6 Population properties for ordinal data

The median estimate of the population is displayed, which is the same value as the sample median.

Additionally, as a hypothesis test, the Wilcoxon signed-rank test is performed to test whether the population median is different from the test value. CogStat The test value is 0 by default, but in CogStat this can be modified before performing the test in the 'Explore variables' dialog in the 'Central tendency test value' field.

Population properties
Median: 3.0



Hypothesis test: Testing if median deviates from the value 0.
Ordinal variable. >> Running Wilcoxon signed-rank test.
Result of Wilcoxon signed-rank test: $T = 0$, $p < 0.001$

Figure Hypothesis test for an ordinal variable

7.7 Population properties for nominal data

CogStat At the moment no population properties are calculated for nominal data in CogStat.

8 Exploring the relation of a variable pair

When at least two variables are measured for the cases in the sample (e.g., the height and the weight of a person), we might be interested whether two variables are related. In other words, we want to see if taking a value in one of the two variables will change the chance what values the other variable can take.

CogStat In CogStat use the 'Analysis > Explore relation of variable pair' menu to perform such an analysis. In the dialog, choose the two variables you want to investigate, then click OK. See the details of the displayed results in the CogStat documentation.

Exercise In the Analysis > Explore relation of variable pair menu set Before and After to be a selected variable, and run the analysis. Check the results of your analysis according to the description below. After looking at your raw data try to estimate your sample and population properties.

8.1 Exploring the relation of two interval variables

8.1.1 Raw data

If any of the two values is missing in a case, the case will be excluded from the whole analysis.

In CogStat, first, the raw data are displayed in a scatterplot. In a **scatterplot** all cases will be displayed as a separate dot in the space, and the location of the case represents the two values that was measured for that case – the location can be considered as the coordinate in that space. The number of dots is the same as the number of cases in that sample (unless there are cases with the same value pairs, where in CogStat the size of the dot is proportional with the number of cases with that value-pairs, i.e., more cases with the same value pairs will be displayed as larger dots).

The relation of the two variables in a scatterplot can be seen as a systematic pattern in that dot cloud. If there is no relation between the two variables, the dots will be random in the scatterplot. When the sample is small, it is really hard to tell by only looking at the scatterplot whether there is some systematic pattern or not.

When there is a relation between the two variables, it can take a lot of forms (see figure below). For example, in a mathematically simple way, the relation can be linear: The dots of the sample will form a straight line. Or it can be a reversed U pattern: for small x values the y values are also small, then for medium x values the y values are larger, finally for large x values the y values are again small. Another frequent option is that the relation is some curved line, such as the logarithmic or exponential mathematical functions.

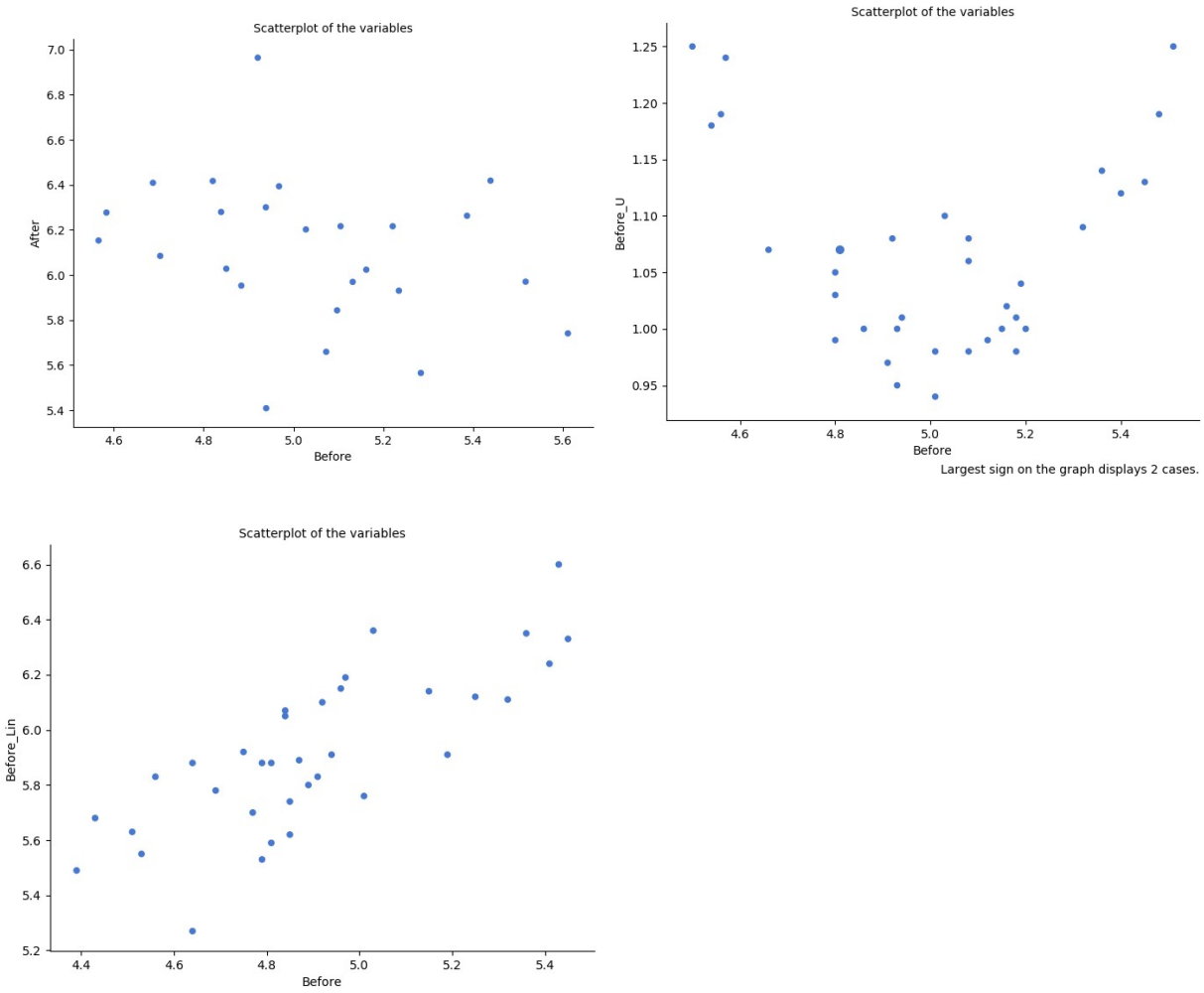


Figure Scatter plots of various relations

Importantly, these relations are almost always noisy: While the dots may fall around an imaginary pattern or line, they do not form a perfect pattern creating clear lines, but they will be more or less random around these imaginary lines.

8.1.2 Linear regression and correlation

Regression analysis. There are various ways to characterize these relations mathematically. Usually, a specific form of a line is chosen – mathematically, this is a function, such as linear function, logarithmic function, power function, etc. Then the parameters of that function are found (i.e., some properties of that line can be changed freely, such as the slope of a linear line) and it is measured how precisely the data follow that function with that parameter. The mathematical term for this procedure is the regression analysis: in a **regression analysis**, a specific function is fit to the data, and the parameters of that fitted function is found, and several mathematical properties of the data related to the fitted function are described.

Linear function. Most frequently, the **linear function** is used for regressions, i.e., in a scatterplot the function would form a straight line. There are various reasons why the linear function is the most popular choice. First, mathematically, the linear function is relatively simple. Second, many times the data are so noisy that linear function seems to be a quite neutral, still appropriate choice.

Parameters of the linear function and linear regression. In a linear function, there are two parameters that determine the position of the line in a chart: The slope and the intercept. The **slope** is the steepness of the line, while the **intercept** is the point where the line crosses the y axes. In the **linear regression**, the two values for these two properties of the line should be found to which linear line the data fit the best. The goodness of the fit can be specified in several ways, and the most frequent method in regression is the **least squares method**. The idea is pretty simple. Imagine a potential fitted line (i.e., choose a specific value for both parameters: for the slope and for the intercept), and draw vertical lines from all dots to the fitted line. Now create a square from all of these vertical lines, i.e., make squares where the size of the side is the length of these vertical lines. Finally, sum the areas of those squares. Exercise The same procedure could be repeated with various fitted lines (i.e., with other slopes and intercepts), and the best fit will be where the sum of the squares is the smallest. Find an interactive demonstration of this method here: Drag the two purple circles to find the line where the sum of squares is the smallest possible value. You might also change the data points by dragging them.

Now checking the sum of squares for various fitted lines sounds like a lot of calculations, but first, usually it is done by the computer, and second, the computer does not have to calculate all possible fitted lines because there is a relatively simple formula that can return the slope and intercept of the fitted line with the smallest sum of the squares.

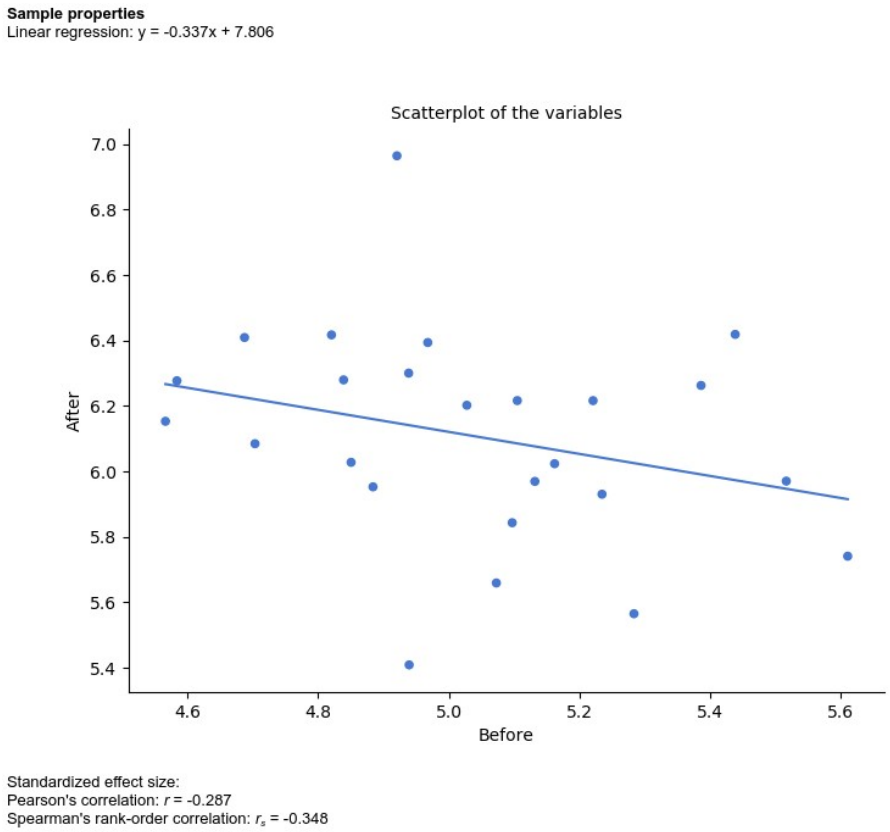


Figure Scatterplot of two interval variables with linear regression results

Overall, in a linear regression analysis, the slope and intercept of the best linear fit can be found, while the best fit is defined by the least square method. CogStat In CogStat, the formula can be found in a form of y = *a* x + *b*, where *a* and *b* are the slope and the intercept, respectively. This line can also be seen in the related chart.

The formula of this linear fit can also be considered as the best guess for a y value if the x is known, supposed that the relation of the two variables is linear. In other words, if you know the x value for a case, then the best guess for the expected y value is to multiply the x value by *a* (i.e., the slope), and add *b* (i.e., the intercept) to it. For example, if the formula shows y = 6 x + 4, and the x value is 3, then the expected y is 6×3+4, which is 22. Obviously, our guess could be imprecise, depending on how closely the dots are around this line.

Goodness of fit. The next thing one might want to know is how strongly that sample fits that line. One way to quantify it is to calculate the **Pearson's correlation coefficient**, or frequently simply called as correlation. The correlation is an index that could take any value between -1 and 1 (see figure below). If the value is 0, it means that the data are entirely random, and they do not fit the line at all. If the value is 1 or -1, it means that the data perfectly fit that line, i.e., all dots are on the line. The difference between the +1 and -1 is that +1 can be seen when the slope of the line is positive, and -1 can be seen when the slope is negative. All other values mean a fit in between the entirely random fit and the perfect fit, where the sign of the correlation is related to the sign of the slope of the line.

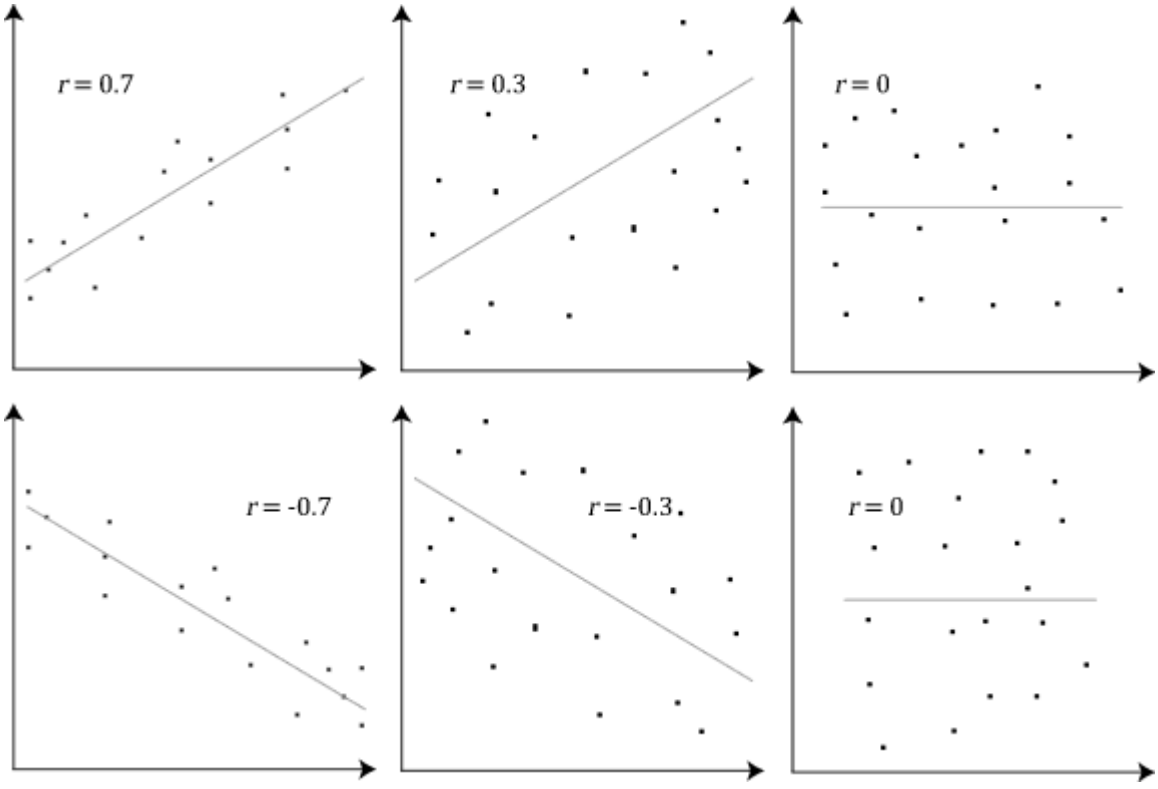


Figure Scatter plots with data of various correlations (source)

Another way to quantify the correlation is the **Spearman's rank order correlation**. This is the correlation of the rank data. Imagine that not the original values but the order of them are displayed in a scatterplot. Now if you try to find the regression line on those rank data, and calculate the Pearson' correlation coefficient, that will be the Spearman's correlation. So the Spearman's correlation is simply a Pearson's correlation on the rank data.

8.1.2.1 Population estimations and hypothesis tests

In CogStat, the population properties section presents the point and interval estimate of the Pearson's and Spearman's correlation coefficient. In the hypothesis test part, it is tested whether the two correlation coefficients deviate from 0. Like in the case of the means in the single variable exploration part, the confidence interval and the hypothesis test results are related: if the test value (here, 0) is outside of the CI, the test is significant, while if the test value is inside the CI, the test is not significant.

Population properties
Standardized effect size:

	Point estimation	95% confidence interval
Pearson's correlation, *r*	-0.287	[-0.613, 0.122]
Spearman's rank-order correlation, r_s	-0.348	[-0.653, 0.055]

Hypothesis test: Testing if correlation differs from 0.
Interval variables. >> Running Pearson's and Spearman's correlation.
Pearson's correlation: $r(23) = -0.287$, $p = 0.164$
Spearman's rank-order correlation: $r_s(23) = -0.348$, $p = 0.089$

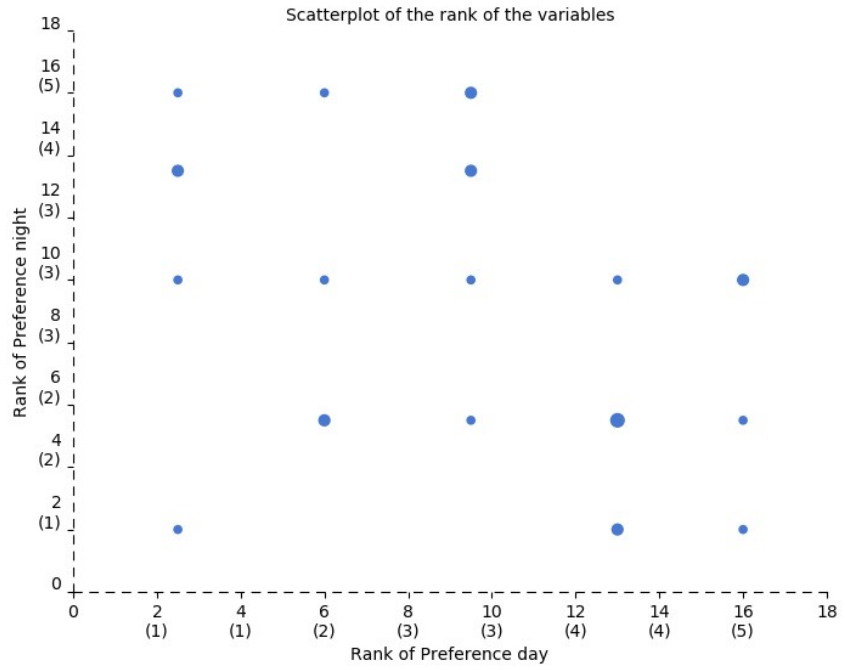
Figure Hypothesis test about the relation of two interval variables

8.2 Exploring the relation of two ordinal variables

The results are similar to the results found for interval variables, with a few differences. First, the scatterplot is based on the order, and not on the value of the data. Second, only rank based Spearman's correlation is calculated, but not the Pearson's correlation – the latter wouldn't make sense for an ordinal variable pair.

Raw data

N of valid pairs: 25
N of missing pairs: 0



Sample properties

Standardized effect size:
Spearman's rank-order correlation: r_s = -0.396

Population properties
Standardized effect size:

	Point estimation	95% confidence interval
Spearman's rank-order correlation, r_s	-0.396	[-0.684, -0.001]

Hypothesis test: Testing if correlation differs from 0.
Ordinal variables. >> Running Spearman's correlation.
Spearman's rank-order correlation: $r_s(23) = -0.396$, $p = 0.050$

Figure Scatterplot of two ordinal variables and the related sample properties, parameter estimations and hypothesis test

8.3 Exploring the relation of two nominal variables

When the relation of two nominal variables is explored, the values of the variables do not have any order, so the only relevant information that creates the raw data is that how many times the specific combinations of the values from the two variables can be observed. To offer this information, the contingency table is calculated: **Contingency table** includes the values of the variables as the rows and columns, and the cells include the count values, how many times a specific value combination occurs. The same information can be seen graphically in the mosaic plot: In the **mosaic plot** the squares denote the combinations of the values of the two variables, and the sizes of the squares are proportional with the number of the cases in those combinations.

The strength of the association between the two variables is measured with the **Cramér's V**. Similar to the correlation coefficients, its value could be between 0 and 1 (although unlike the correlation coefficients, this value cannot be a negative number), where 0 means no relation between the variables, and 1 means maximal connection (i.e., if one knows the value of one of the variables, one can tell the value of the other variable) between the variables. Finally, as a hypothesis test, a Chi-square test is performed to find out whether the relation is different from 0.

Raw data

N of valid pairs: 25
N of missing pairs: 0

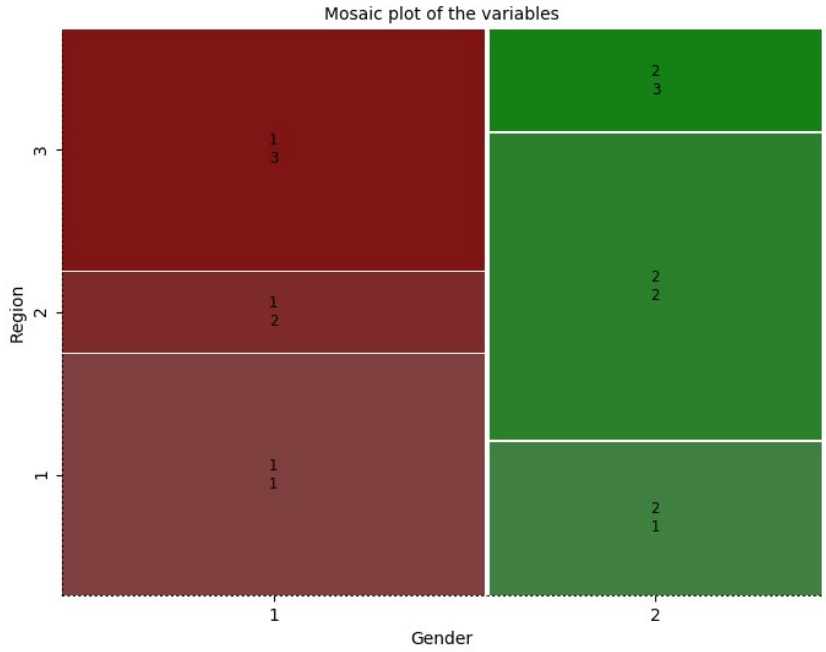


Figure Mosaic plot of two nominal variables

9 Comparing groups

In CogStat, to compare the values in a variable across groups, choose the ‘Analysis > Compare groups’ menu, and choose the dependent and grouping variables. See the details of the displayed results in the CogStat documentation.

In this comparison, the grouping variable is always considered as a nominal variable, even if it is an ordinal or interval variable. In other words, in the grouping variable no order or specific values are considered (e.g., schooling, salary categories, etc.), they are just considered as different groups ignoring all other possible information in that variable.

Exercise In the Analysis > Compare groups menu set Task_performance to be the dependent variable, set Region to be a grouping variable, and run the analysis. Check the results of your analysis according to the description below. After looking at your raw data try to estimate your sample and population properties.

9.1 Comparing groups with interval dependent variables

First, the number of missing data are displayed. There could be two reasons why a data is missing: either dependent variable value can be missing, or the grouping variable value can be missing. Cases with missing data will be excluded from the analysis.

In the raw data, the single cases can be seen in a dot plot, like when exploring a single variable, but here the data are grouped according to the grouping variable (i.e., groups are displayed in separate columns).

Raw data

	1	2
N of valid cases	12	13
N of missing cases	0	0

N of missing group cases: 0

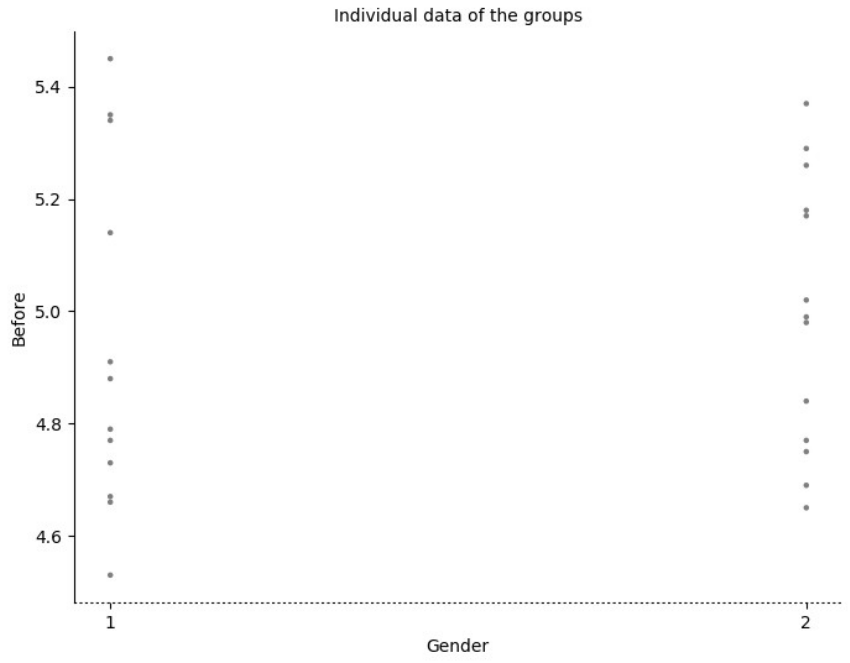


Figure Raw data of two groups of interval variables

The sample properties include a few descriptives for the groups, such as the means, standard deviations, and so on, and it displays the box plot of the data per groups together with the individual data, just like when a single variable is explored, but now split into groups.

Sample properties
Descriptives for the groups

	1	2
Mean	4.935	4.997
Standard deviation	0.295	0.234
Maximum	5.450	5.370
Upper quartile	5.190	5.180
Median	4.835	4.990
Lower quartile	4.715	4.770
Minimum	4.530	4.650

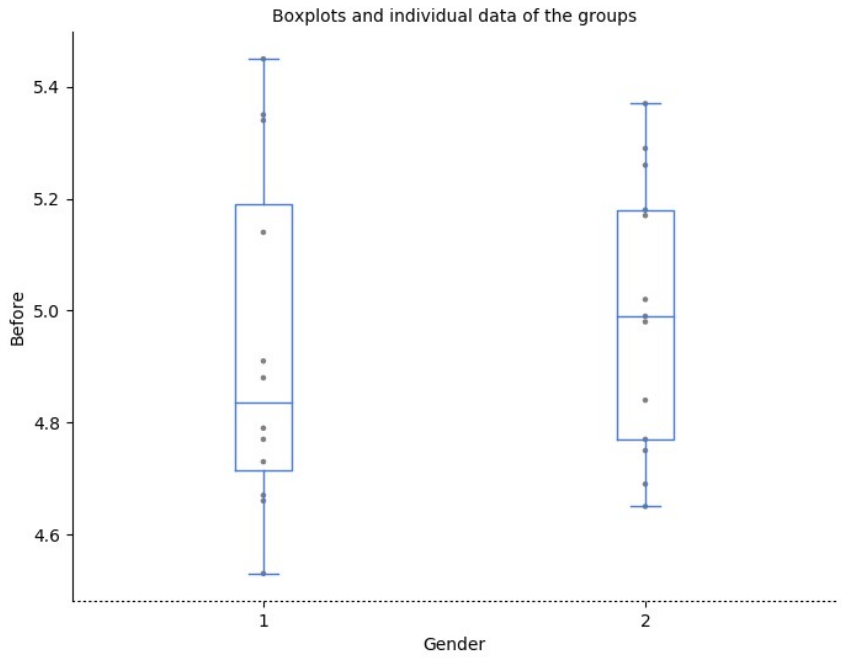


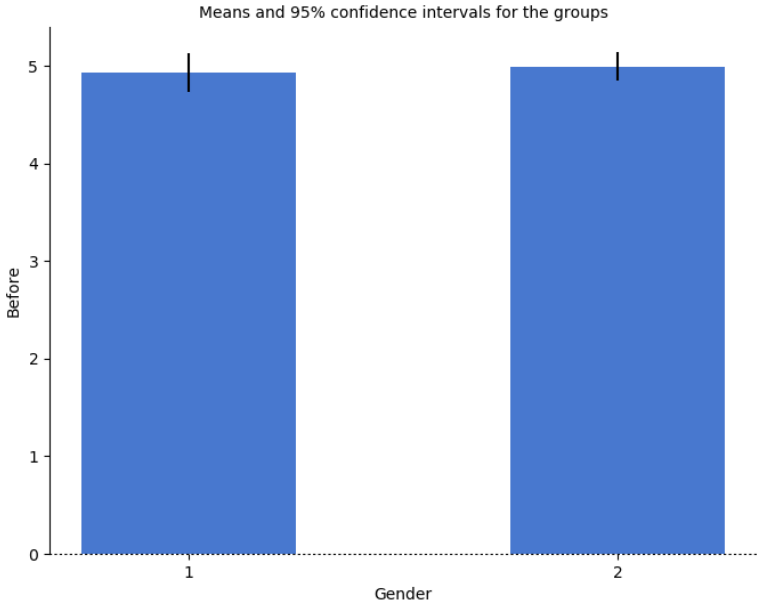
Figure Descriptive results of two groups of interval variables

In the population properties, the point estimation of the means and the confidence intervals of the means are displayed. In the hypothesis test section, the difference of the means is tested. After the appropriate assumptions are checked (see the details in the CogStat documentation), the hypothesis tests are run. See the documentation of CogStat what hypothesis tests are used and what the choice depends on.

Population properties

Means
Present confidence interval values suppose normality.

Gender	Point estimation	95% CI (low)	95% CI (high)
1	4.935000	4.738951	5.131049
2	4.996923	4.849905	5.143941



Hypothesis testing: Testing if the means are the same.
One grouping variable. Two groups. Interval variable. >> Choosing two sample t-test, Mann-Whitney test or Welch's t-test depending on assumptions.
Checking for normality.
Shapiro-Wilk normality test in variable Before (Gender: 1): $W = 0.9$, $p = 0.160$
Shapiro-Wilk normality test in variable Before (Gender: 2): $W = 0.939$, $p = 0.447$
Checking for homogeneity of variance across groups.
Levene test: $W = 0.404$, $p = 0.531$
Normality and homogeneity of variance are not violated. >> Running two sample t-test.
Difference between the two groups: -0.062, 95% confidence interval [-0.291, 0.167]
Result of independent samples t-test: $t(23) = -0.56$, $p = 0.581$

Figure Parameter estimations and hypothesis tests of two groups of interval variables

9.1.1 Post hoc tests

When more than two groups are compared, a significant hypothesis test result about the group means denotes that among the several groups the mean of at least one of those groups differs from the mean of at least another one of those groups. This significant test will not tell whether there is a single pair of groups differing or there are several pairs. Also, the significant test will not tell in which specific group pairs differ the means. To find the specific group pairs that cause the significant group difference in the first omnibus test, post hoc tests are performed. These post hoc tests will test the groups pairwise whether they differ from each other or not. (CogStat will run the post hoc tests automatically if the omnibus test is significant.)

9.2 Comparing groups with ordinal dependent variables

When groups with ordinal dependent variables are compared, very similar results will be shown as for the interval variables, with a few differences. In the raw chart and in the box plot, the order of the data is displayed instead of the values of the data. In the sample properties, only the median and some other ordinal descriptives are displayed. Finally, the hypothesis test is much simpler

because no assumptions should be checked for those tests. Find the specific hypothesis test CogStat chooses in the CogStat documentation.

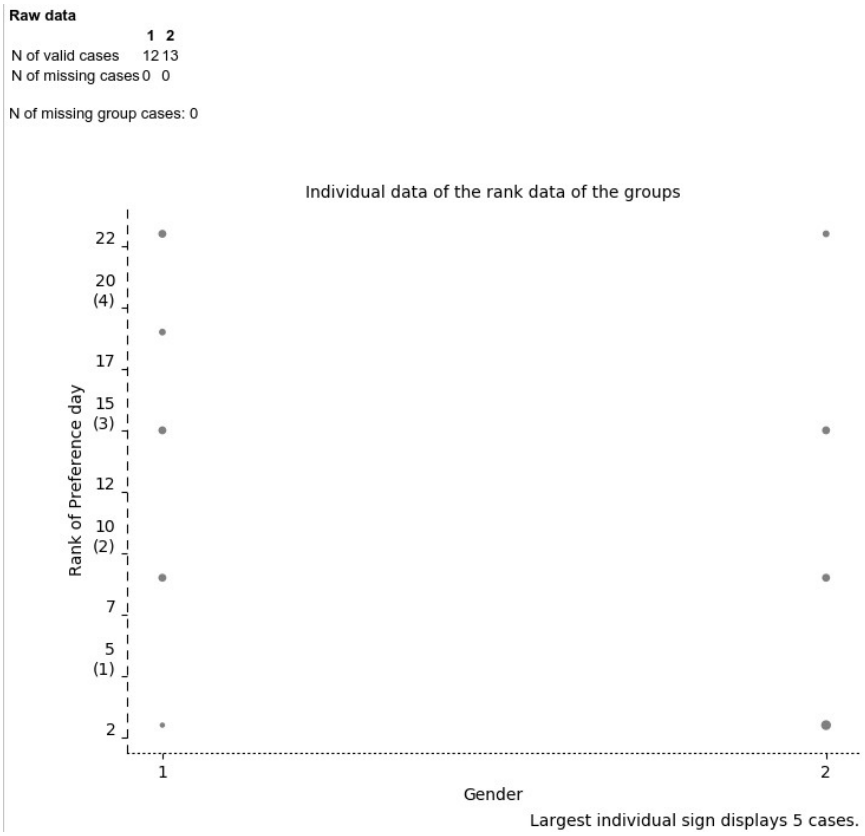
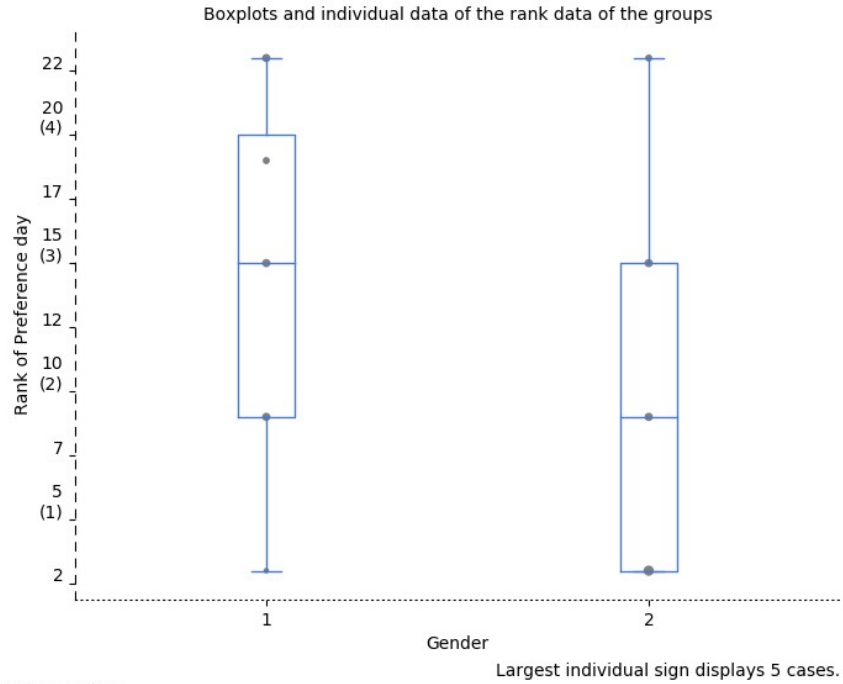


Figure Raw data of two groups of ordinal variables

Sample properties
Descriptives for the groups

	1	2
Maximum	5.0	5.0
Upper quartile	4.2	3.0
Median	3.0	2.0
Lower quartile	2.0	1.0
Minimum	1.0	1.0



Population properties

Hypothesis testing: Testing if the medians are the same.
One grouping variable. Two groups. Ordinal variable. >> Running Mann-Whitney test.
Result of independent samples Mann-Whitney rank test: $U = 108$, $p = 0.094$

Figure Descriptive results, parameter estimations, and hypothesis tests of two groups of ordinal variables

9.3 Comparing groups with nominal dependent variables

Because both the grouping and the dependent variables are nominal, the question whether a nominal grouping variable has an effect on another nominal variable, is in fact the same as the question about the relation of two variables. For this reason, practically the same results are calculated as when exploring the relation of two nominal variables.

10 Comparing repeated measures variables

When comparing groups, one is interested in the effect of the groups: How group membership influences the dependent variable. In such a study, the researcher might influence which case should belong to which group, for example, investigating the effect of text color on reading speed, the experimenter can specify which participants belong to the red text group, and which participants belong to the green text group. (Note that some factors cannot be investigated in this way. For example, when we want to compare men and women, that property is already given for the participants, and the researcher cannot categorize arbitrarily the participants, saying for example, that this participant will be man in this study, but woman in the next one.) However, the effect of some factors could be investigated with a different method as, well. In the color text reading example, all participants could read both the red and the green color texts, and we could measure the reading speed in both conditions. This latter study design is termed **within subject design** (the conditions, e.g., the colors, are compared within the participants) in the methodology literature, while the group-based design is termed **between subject design** (where the conditions are compared between the participants, creating groups of participants). In the statistical terminology, within subject design is termed repeated measures because the same thing is measured repeatedly (i.e., in different condition) for a case (here, for participants).

When the same thing is measured repeatedly, in an analysis we want to compare the different measurements, or in other words, we want to compare the different repeated measures variables, and in the methodological terminology we want to compare the different conditions.

From a mathematical viewpoint, the main difference between the within subject design and the between subject design is that in repeated measurements a single participant is measured in all conditions, while in the between subject design there is no such relation between the conditions, but all values come from different participants. The analysis can take advantage of the fact that a single participant is measured in all conditions. In the between subject design, the groups may differ not only because the grouping variable has an effect, but because all groups include different participants, which will introduce some noise (noise in a sense that these individual differences are not controlled in the study). On the contrary, in the within subject design, this source of noise is excluded because all values in a condition have their related values in other conditions which related values come from the same participant. The overall consequence of this difference is that the within subject design excludes a noise source: Individual differences between the conditions. Therefore, the within subject design data can reveal phenomena more efficiently than the between subject design data.

CogStat In CogStat, to run a repeated measures comparison select the ‘Analysis > Compare repeated measures variables’ menu, and select the variables that include the different measurement points. See the details of the displayed results in the CogStat documentation.

Exercise In the Analysis > Compare repeated measures comparison menu set Before and After to be selected variables, and run the analysis. Check the results of your analysis according to the description below. After looking at your raw data try to estimate your sample and population properties.

10.1 Comparing repeated measures interval variables

The result for a comparison of repeated measures interval variable is very similar to the comparison of groups with interval variables. To highlight the critical information that the data of the different conditions are measured for the same case (e.g., the same participants performs the task repeatedly in different conditions), the data of a single case are connected with a line across the conditions when the raw data are displayed. Another difference between repeated measures comparison and group comparison is that the specific hypothesis tests are different in repeated measures and group comparison analysis (see more details in the CogStat documentation).

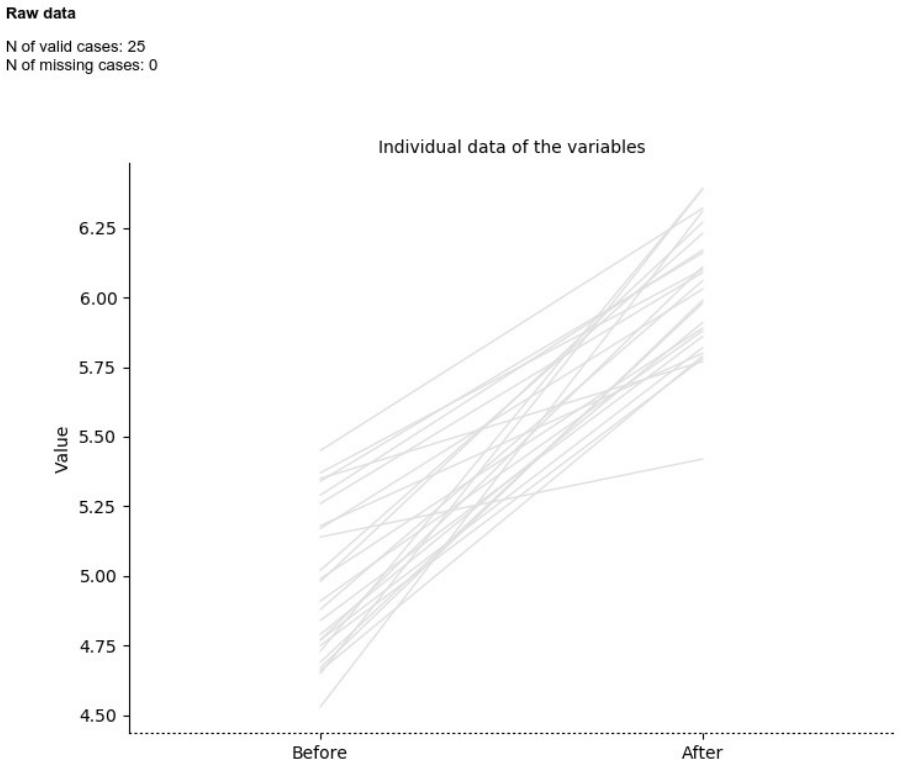


Figure Raw data of two repeated measures conditions of interval variables

Sample properties
Descriptives for the variables

	Before	After
Mean	4.967	6.021
Standard deviation	0.267	0.229
Maximum	5.450	6.390
Upper quartile	5.180	6.170
Median	4.910	6.030
Lower quartile	4.750	5.860
Minimum	4.530	5.420

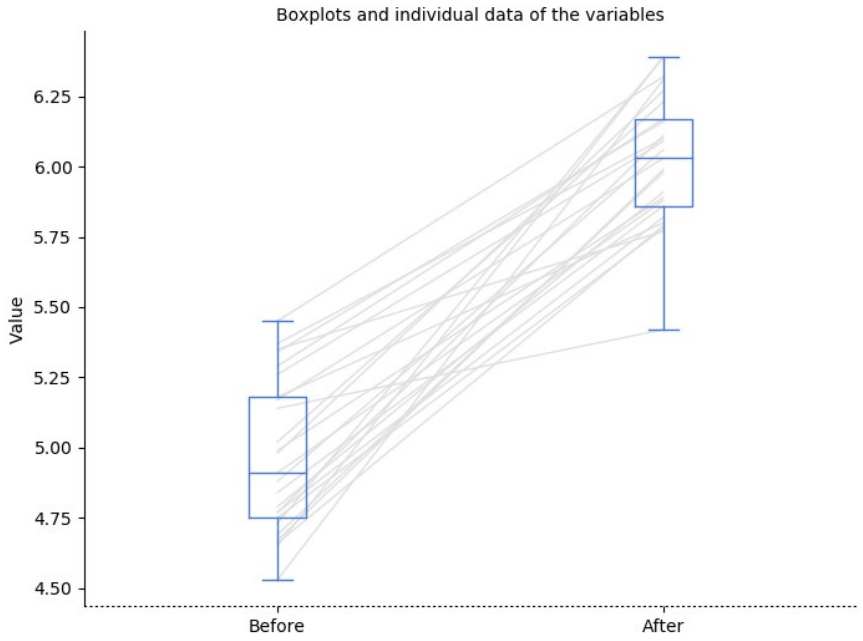
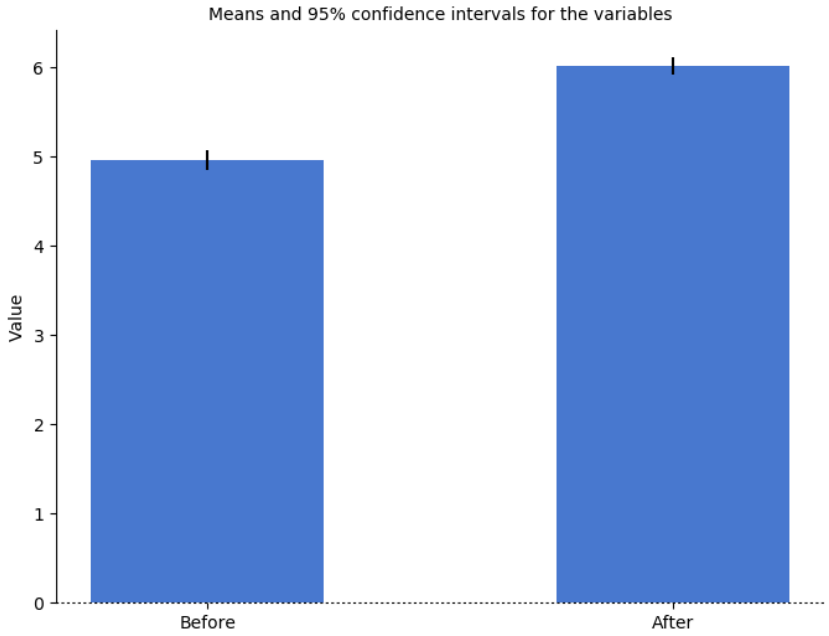


Figure Descriptive results of two repeated measures conditions of interval variables

Population properties

Means
Present confidence interval values suppose normality.

	Point estimation	95% CI (low)	95% CI (high)
Before	4.9672	4.854738	5.079662
After	6.0208	5.924149	6.117451



Hypothesis testing: Testing if the means are the same.
Two variables. Interval variables. >> Choosing paired t-test or paired Wilcoxon test depending on the assumptions.
Checking for normality.
Shapiro-Wilk normality test in variable Difference of Before and After: $W = 0.972$, $p = 0.704$
Normality is not violated. >> Running paired t-test.
Result of paired samples t-test: $t(24) = -14.4$, $p < 0.001$

Figure Parameter estimation and hypothesis test of two repeated measures conditions of interval variables

10.2 Comparing repeated measures ordinal variables

Results of ordinal variables are very similar to results for interval variables. The only difference at the moment is that the hypothesis tests differ (see more details in the CogStat documentation. Note that unlike in other parts of CogStat, at the moment, for an ordinal variable it is not the order that is displayed in a chart, but the original values; this is intended to be changed in a later release of CogStat.

10.3 Comparing repeated measures nominal variables

When checking whether the same values were measured in repeated occasions, one can investigate the value combinations between two occasions, resulting in a similar contingency table and mosaic plot as in the relation of two nominal variables.

Still, conceptually this is not exactly the same thing as the relation of two variables or the similar comparison of groups with nominal dependent variable. The main difference is that for both the relation analysis and the group comparison the nominal variables could be any variables, the only

criteria is that they should be nominal. On the other hand, in a repeated measures comparison, the variables should be not only nominal, but they should measure the same thing.